TABLE OF CONTENTS

		Page	
LIST OF FIGURES		vii	1/A5
LIST OF TABLES		viii	1/A6
LIST OF SYMBOLS		x	1/A8
1. INTRODUCTION		1	1/A13
	a. Statement of the problem	1	1/A13
	b. Previous studies	2	1/A14
	c. Objectives	7	1/B5
	d. Importance	9	1/B7
2. STATISTICAL APPROACH		10	1/B8
	a. Linear models	10	1/B8
	b. Partitioning sums of squares	11	1/B9
	c. Model assumptions and violations	13	1/B11
	d. Multicollinearity	19	1/C3
	e. Variable selection methods and inference	22	1/C6
	f. Interpretation of regression model results	24	1/C8
	g. Principal component analysis	28	1/C12
3. DATA SELECTION AND PROCESSING		31	1/D1
	a. Location	31	1/D1
	b. Surface data	31	1/D1

TABLE OF CONTENTS (CONTINUED)

		Page	
c.	Upper-air data	31	1/D1
d.	Radar data	31	1/D1
e.	Initial processing	35	1/D5
f.	Objective analysis	36	1/D6
4.	PARAMETERIZATION AND DATA SUBDIVISION	38	1/D8
a.	Predictand formulation	38	1/D8
b.	Predictor formulation	38	1/D8
c.	Subdivisions of original data	48	1/E4
5.	RESULTS	52	1/E8
a.	Forecast time intervals	52	1/E8
b.	Predictor selections	53	1/E9
c.	Importance of surface versus upper-air parameters . .	56	1/E12
d.	Quality of fit of the regression model	60	1/F2
e.	Performance on a test data sample	64	1/F6
f.	Comparison with other results	70	1/F12
g.	Dimensionality	74	1/G2
h.	Operational utility	78	1/G6
6.	UPPER-AIR CONDITIONS AT 3-h INTERVALS	82	1/G10
7.	SUMMARY AND CONCLUSIONS	91	2/A14

TABLE OF CONTENTS (Concluded)

	Page	
8. SUGGESTIONS FOR FURTHER RESEARCH	94	2/B3
REFERENCES	97	2/B6
APPENDICES	103	2/B12

LIST OF FIGURES

Figure		Page	
1	Plot of mixing ratio versus residual from regression model	16	1/B14
2	Plot of residuals versus predicted values from regression model	18	1/C2
3	A logistic response function	20	1/C4
4	Data reporting locations	32	1/D2
5	Manually digitized radar grid	34	1/D4
6	Orientation of the predictand area with respect to the predictor point	39	1/D9
7	Schematic illustration of isoline concentration by (a) shear and (b) confluence	45	1/E1
8	Subdivision of total data set	49	1/E5
9	Relation between predicted and observed probabilities of thunderstorm occurrences for various data subdivisions . .	63	1/F5
10	Regression lines for data in Tables 3 and 4	69	1/F11
11	Frequency distributions of occurrence and nonoccurrence of MDR≥4 for different values of the surface mixing ratio . .	71	1/F13
12	Fractional amount of total variance in thunderstorm occurrence accounted for by numbers of predictors and a combination of selection procedures	83	1/G11
13	MDR data for the AVE IV experiment	85	1/G13
14	Synoptic charts for 2100 GMT, 24 April 1975 (Fucik and Turner, 1975)	87	2/A10

LIST OF TABLES

Table		Page	
1	Analysis of variance	12	1/B10
2	Data and regression analysis for 10 observations of hypothetical variables x and y with 10% occurrence frequency	26	1/C10
3	Data and regression analysis for 10 observations of hypothetical variables x and y with 30% occurrence frequency	27	1/C11
4	Data and regression analysis for 57% of nonoccurrence observations in Table 3 data	27	1/C11
5	Comparison of Tables 2, 3, and 4	28	1/C12
6	Explanation of Manually Digitized Radar (MDR) code . . .	33	1/D3
7	Summary of analysis parameters	37	1/D7
8	Candidate predictors	41	1/D11
9	Variables selected, cumulative R^2, and sign of coefficient for a stepwise selection procedure and different data subsets	54	1/E10
10	Linear correlation coefficients of selected predictors with the occurrence of thunderstorms during the period 2000-2300 GMT	57	1/E13
11	Summary of statistics for regression analyses with surface and upper-air predictors	59	1/F1
12	Summary of statistics for regression analyses with different data subsets	61	1/F3
13	Contingency tables for 25 no-thunderstorm forecasts shifted to the yes forecast column in different observed proportions	65	1/F7
14	Summary of contingency tables	67	1/F9
15	Contingency table of observed and forecast thunderstorms for 14 base weather stations near the area outlined in Fig. 4	73	1/G1

LIST OF TABLES (Concluded)

Table		Page	
16	Eigenvectors and eigenvalues for moisture, stability and trigger parameters	76	1/G4
17	Eigenvalues and cumulative portion of total variance accounted for by each successive eigenvector	79	1/G7
18	Summary of AVE IV results for stepwise (A), maximum R^2 reduction (B), and stepdown (C) variable selection techniques	88	2/A11

LIST OF SYMBOLS

A_8, A_7, A_5, A_3	subscript convention; A is any variable; A_8, A_7, A_5, A_3 represent the value of A at the 850, 700, 500, and 300 mb pressure levels, respectively
$\lvert\vec{A}\rvert$	magnitude of a two-dimensional vector quantity $\sqrt{A_x^2 + A_y^2}$
ALTSTG	altimeter setting
C_p	Specific heat at constant pressure (1.00488×10^3 J $kg^{-1}K^{-1}$)
CSIL	convective stability index, low level
CSIM	convective stability index, mid level
$\lvert\vec{D}\rvert$	magnitude of a discontinuity function (see text)
DTA	differential temperature advection
DTH	differential thickness
DVA	differential vorticity advection
e	vapor pressure
F	a test statistic for a regression analysis
g	the acceleration of gravity (9.8 m s^{-2})
IDIV	integrated divergence 850 to 500 mb
IMDIV	integrated moisture divergence 850 to 300 mb
KI	K stability index
L	latent heat of condensation (assumed constant at 2.5122×10^6 J kg^{-1})
m	either the number of predictors or when used as a subscript, the mth predictor
MDIV	moisture divergence
MDR	manually digitized radar data
MDRP	manually digitized radar predictor

MSE	mean square error
MSR	mean square due to regression
n	total number of observations
P	sea level pressure
p	probability of occurrence
PY	predicted value of Y
R	gas constant for dry air (2.8704×10^2 J $kg^{-1}K^{-1}$)
R^2	ratio of SSR/SST; also called coefficient of variation
SS	skill score
SSE	sum of squares of errors in a regression analysis of variance
SSR	sum of squares due to regression
SST	total corrected sum of squares
STSI	static stability index
T	temperature
T_d	dew point temperature
TS	threat score
TTI	Total Totals Stability Index
u	East-West wind component
UWSH	850 to 500 mb wind shear of the u-wind components
$\vec{V}$	horizontal wind vector
v	North-South wind component
$\lvert\vec{V}_5\rvert$	magnitude of 500 mb wind speed
VSUM	sum of v-wind components at 850 and 500 mb

LIST OF SYMBOLS (Concluded)

W	mixing ratio
W_s	saturated mixing ratio
x_{im}	the ith observation for the mth candidate predictor
Y	the dependent variable; the occurrence (one) or nonoccurrence (zero) of thunderstorms
Z	geopotential height
β	partial regression coefficient
$\vec{\nabla}$	two dimensional del operator with vertical component excluded
$\vec{\nabla}^2$	horizontal Laplacian operator $(\partial^2/\partial x^2 + \partial^2/\partial y^2)$
$\vec{\nabla}_p$	two-dimensional del operator on a constant pressure surface
$\vec{\nabla}^2_{THA}$	Laplacian of thickness advection
ε	residual or error term in a regression model
θ	potential temperature
θ_e	equivalent potential temperature
$\theta_e A$	advection of equivalent potential temperature
ζ	vertical component of wind curl (vorticity)
ρ	air density
ω_T	terrain induced vertical motion
ω_{TS}	vertical motion at the top of the surface layer (see text)

DEC 1 8 1977

COMPLETED

ORIGINAL

NASA Contractor Report 2934

Forecasting Thunderstorms Over a 2- to 5-h Period by Statistical Methods

Joseph Allen Zak

CONTRACT NAS8-31773
DECEMBER 1977

NASA

NAS 1.26:2934

NASA Contractor Report 2934

Forecasting Thunderstorms Over a 2- to 5-h Period by Statistical Methods

Joseph Allen Zak
Texas A&M University
College Station, Texas

Prepared for
George C. Marshall Space Flight Center
under Contract NAS8-31773

National Aeronautics
and Space Administration

Scientific and Technical
Information Office

1977

AUTHOR'S ACKNOWLEDGMENTS

The author wishes to express his sincere appreciation to Dr. James R. Scoggins for his guidance throughout this project. Additional thanks are due to Dr. Alymer Thompson, Dr. Dennis Driscoll, Dr. Jack Barnes, Dr. William Perry, and Dr. Harry Coyle for their assistance during the final preparation of this report. He is also grateful to Dr. Rudolf Freund for many helpful suggestions concerning the statistical treatment of the data and interpretation of the results. Thanks are due to many personnel of NOAA's Techniques Development Laboratory for sharing their data and expertise at the beginning of this project and to Mr. Gregory S. Wilson for programming assistance.

This research, which was submitted initially to the Graduate College of Texas A&M University in partial fulfillment of the requirements for the degree of Doctor of Philosophy, was sponsored in part by the National Aeronautics and Space Administration under Contract NAS8-31773 under the auspices of the Atmospheric Sciences Division, Space Sciences Laboratory, NASA/Marshall Space Flight Center, Alabama, and by the Department of Meteorology, Texas A&M University. None of this research would have been possible without a scholarship from the Air Force Institute of Technology, which enabled the author to attend Texas A&M University.

Finally, special thanks are due to Miss Karen Cobbs for typing the manuscript and to Miss Doreen Westwood for preparing the figures.

1. INTRODUCTION

a. Statement of the problem

Thunderstorms are meteorological phenomena of great importance to meteorologists because of the energy conversions and momentum transports which occur. Manifestations of the above are the damaging winds and hail so often observed. Unfortunately, the prediction of the occurrence and intensity of these storms has been a problem of substantial significance for meteorologists that has defied easy solution. There are several reasons for this. First, a thunderstorm ranges in diameter from a few tens to one hundred kilometers and lasts on the order of 10^3 to 10^4 seconds. Such mesoscale phenomena elude detection by most routine observations. Also, the analysis and forecast schemes that are in operational use are applied to areas and time scales much greater than these. The larger scales permit only a degree of success in predicting large areas in which the likelihood of thunderstorm occurrence is great (Fawcett, 1977). Another reason is that our knowledge of the dynamics and thermodynamics of thunderstorms is not sufficient to explain these phenomena. Also, the precise nature of the interactions between the large- and small-scale circulations is not sufficiently well understood (Barnes, 1976) for the purpose of exact forecasting.

One approach to the solution of the forecasting problem is through parameterization of large-scale processes and use of appropriate statistical techniques. There may be information from present observations that, when used in certain combinations, can improve the prediction of

---

The citations on the following pages follow the style of the Journal of Applied Meteorology.

thunderstorms over a 2- to 5-h period. For periods less than 2 hours, persistence and radar pattern recognition techniques should give the best results. For periods beyond 5 h, it is unlikely that observations will reflect the structure of the atmosphere which produces thunderstorms. Furthermore, there may be improvement in prediction if upper-air data were collected at more frequent intervals. Finally, optimum combinations of parameters determined by statistical techniques may lead to improved physical models.

The hypotheses underlying this research are that manifestations of the thermodynamic and hydrodynamic interactions which evolve into intense convection in the atmosphere can be detected in routine observations and that these observed parameters can be used in a statistical model (which minimizes the unexplained variance of observed thunderstorms) for prediction.

b. Previous studies

1) Nature of thunderstorms

Thunderstorms occur in comparatively small regions in the atmosphere. Prior to 1947 there were few measurements of meteorological variables in and near thunderstorms, so that circulation, pressure, temperature, and moisture patterns were known only qualitatively. With the realization of the Thunderstorm Project (Byers and Braham, 1949), however, our quantitative knowledge increased significantly. Measurements collected over a 2-yr period established the horizontal and vertical structure of many meteorological variables associated with thunderstorms and confirmed the existence of multiple convective cells in various stages of development.

Scorer and Ludlam (1953) proposed a bubble theory of convection that explains many of the observed features of a growing convective element. In this concept, the kinematics resemble those of a spherical vortex, as discussed by Woodward (1959) and Turner (1964). Later stages better resemble a jet of upward-moving air (Squires and Turner, 1962) which exists in nearly steady state, particularly in the presence of vertical wind shear. Ludlam (1963) discussed the role of the tilted updraft core, a manifestation of wind shear, as a natural way to shield the updraft that generates energy from the destructive influences of precipitation-induced downdrafts and environmental entrainment.

Recent meteorological literature contains many articles concerning thunderstorms, their interactions, intensification, movement, and structure. It is not our purpose to review these in detail, but the following synopsis will point out the complexity of thunderstorms and environmental interactions with which we must be concerned.

Thunderstorms grow from a few kilometers in diameter to large, quasi-steady supercells 20-50 km in diameter (Browning and Ludlam, 1962). They can last from 30 min to many hours. Such storms may or may not spawn tornadoes, rotate, contain destructive downdrafts or hail, or exist in strong wind shear. Even the simple cumulus source is not simple at all, as pointed out by Auer (1976) from his observations of distortions of θ_e fields near a cloud boundary. The entraining plume model falls short of describing the thunderstorm documented by Saunders and Paine (1975). In this severe supercell there was little downdraft at the surface, but a mesoscale updraft-downdraft doublet aloft seemed to permit vertical motions to persist for several hours without large

perturbations in isentropic surfaces. Lemon (1976) discusses a flanking line thunderstorm which includes both multicell and supercell storms that derive impetus from entrainment of flanking cells. Still another category termed "spearhead echo" by Fujita and Byers (1977) has intense destructive downdrafts which appear to be tied to overshooting tops of clouds at the anvil level. Finally, a fascinating observation that "the growth of vigorous squall lines and severe weather are sharply inhibited at and to the south of the subtropical jet" is documented and explained by Whitney (1977).

As new mesoscale observational tools, such as Doppler radar and storm satellites (Shenk, et al., 1976), are added to our operational inventory, we are likely to observe even more differences among thunderstorms. Now, we have observations of internal motions within cells from experimental Doppler radar (see, for example, Brandes, 1977; Kropfli and Miller, 1976). Complicated motion patterns of outflow aloft and jet stream interaction can be observed from stationary satellite picture composites. The intricate details of overshooting, which seem to be linked to tornado formation, can be seen from satellite film loops as well.

There is no "typical" thunderstorm. Each storm is unique in many respects. It is highly unlikely that identical environmental impulses exist on different days or even in different locations on the same day. It is not surprising that modelers and forecasters have much difficulty in their tasks of understanding and forecasting these phenomena.

Concerning the environment, we know that conditions necessary for

severe convective development involve a) convective instability and a lifting mechanism to release it, b) abundant low-level moisture over which a dry-air intrusion exists, and c) bands of strong winds in the lower and upper levels (Miller, 1972). For less severe storms this list reduces to moisture, potential instability[1], and a trigger. These conditions must be identified through existing meteorological data networks and numerical prognoses.

2) Forecasting procedures

Present forecasting procedures are somewhat subjective, and therefore strongly influenced by a person's knowledge and experience. As these vary with individuals who tend not to stay at one location, thunderstorm forecasting procedures for a given point are highly variable. A typical forecast involves 1) a study of the existing and past large-scale weather patterns with emphasis on the location of discontinuities and features discussed in the preceding paragraph, 2) an analysis of stability of the atmosphere from the nearest and latest upper-air sounding, 3) evaluating the latest available numerical forecasts and interpolating for a given time and location, 4) a closer look at the local weather and hourly changes, particularly from surface observations and radar, and 5) a decision on whether or not all the ingredients for thunderstorms will exist at the station for the future time in question. This last step requires synthesizing all the data from the previous steps.

Objective techniques offer several advantages. They do not require

---

[1] Defined by Palmén and Newton (1969, p. 345) to include both convective and conditional instability.

extensive personal experience; they can synthesize a great amount of data rapidly and effectively; they can be automated. Furthermore, they can be developed to make maximum use of historical observations. Finally, established rules for parameterizations can be followed.

There are three steps in a parameterization approach. First, one must know the processes (equations) involved. Next, relevant parameters must be combined in an appropriate functional relationship. Finally, one must test the results. A more detailed description of parameterization techniques is given in the Global Atmospheric Research Programme (GARP) Publication No. 8 (1972).

A statistical approach to thunderstorm forecasting is used partly to alleviate the disparity between the lack of understanding and the need for prediction, partly to glean as much information as possible from existing data, and partly to gain the benefits of objective forecasting schemes. The forecasting of mesoscale phenomena by statistical techniques is not new. Persistence probability has aided the operational forecaster in predicting changes in ceiling and visibility as well as the onset and duration of critical values of meteorological variables. Endlich and Mancuso (1968) combined a number of measured atmospheric quantities into several kinematic and thermodynamic parameters which were correlated with severe thunderstorms and tornadoes. Similarly, observational data were used in an objective (statistical) procedure to forecast severe thunderstorms and tornadoes by Miller and David (1971). Probability-of-precipitation forecasts and other model output statistics have been available for several years (Glahn and Lowry, 1972). More recently, 24-h forecasts of probabilities of thunderstorms and severe

thunderstorms have become available from the National Weather Service (Alaka et al., 1973). In these procedures, various potential predictors from numerical forecasts were used in a screening regression program. Those predictors selected account for a certain fraction of the total variance of observed thunderstorms as derived from historical manually digitized radar (MDR) data (Moore et al., 1974). Finally, a statistical regression forecast for severe thunderstorms 2 to 6 h in the future also recently became available (Charba, 1975). General thunderstorm forecasts were added during the spring of 1976, and other improvements were made in 1977 by Charba (1977) (see, also, the National Weather Service Technical Procedures Bulletin 194).

In these latter procedures predictors were derived from surface observations and dynamic model forecasts. An advantage to the use of parameters from forecast models is that the physics of the circulation system is included. A disadvantage, however, is that changes to the model necessitate development of new equations, as the old regression equations apply only to variables calculated from the former model. Another disadvantage is that inaccuracies in the forecasts will limit the degree to which the model can describe the predictand. Finally, predictors lose their simple interpretation in that forecast elements include biases from the model. In this research the disadvantages are eliminated, and the physics will be included to the greatest extent possible in the choice of candidate predictors.

c. Objectives

Within the general framework of developing a statistical model to forecast thunderstorms in a 2- to 5-h period will be the following

objectives: developing parameters for candidate predictors that are consistent with known physical processes, parameterization methods, and interactions between systems of different scale, relating various test statistics to available verifications of existing thunderstorm-forecasting methods, developing a way to use the spatial variation of meteorological variables to best advantage when many independent variables are involved, interpreting statistical results in terms of violations of model assumptions, assessing the influence of upper-air observations available at 3-h intervals, and finding optimum times for the dependent variable and time changes for selected predictors.

This research will extend the work of Charba (1977) and others in several important ways. First, different statistical models will be evaluated such as principal component analysis, variable selection, and discriminant analysis. Analysis-of-variance statistics will be examined along with plots of key parameters to determine the magnitudes of errors due to assumptions made in the models. Secondly, the final model will be tested on an independent data sample. These statistics will be related to actual verifications of thunderstorm forecasts. Thirdly, upper-air observations will be employed and their importance to observed (by radar) thunderstorms assessed. A unique set of upper-air data collected during atmospheric variability experiments (Fucik and Turner, 1975) will permit calculations of upper-air parameters every 3 hours for one day. These data are available usually at 12-h intervals. Finally, potential predictors will be calculated from the observed variables in a way which will minimize intercorrelations which exist naturally in this type of data.

As long as a short-period forecasting requirement exists, meteorologists must strive to produce the best forecasts possible. This research will contribute to that goal, and may also aid in the underlying goal of understanding the complex interactions of atmospheric parameters which culminate in thunderstorms.

d. Importance

Meteorological data networks and numerical forecasting techniques are established for the synoptic scale of atmospheric analysis and prediction. A true mesoscale data network is prohibitively costly and could not be handled with present computer systems. Until new observational tools such as Doppler radar and geosynchronous satellites are perfected and automated, we are constrained in making point forecasts of mesoscale phenomena such as thunderstorms with present-day data. These data consist of 1) hourly surface reports from stations spaced approximately 150 km apart, 2) hourly radar reports manually digitized from a network in the eastern two-thirds of the United States, 3) satellite photographs at 30-min intervals available at selected locations, and 4) 12-h upper-air observations from stations spaced approximately 300 km apart. Our task, then, must be to extract as much information as possible from these data. This is made more realistic, physically, by the postulate that the energy required to initiate the development of mesoscale systems is contained within the synoptic-scale systems (Global Atmospheric Research Programme, 1972, p. 1).

2. STATISTICAL APPROACH

The theory of classical statistical methods such as least squares and regression analyses is well documented (see, for example, Draper and Smith, 1966; Morrison, 1976; Neter and Wasserman, 1974), and will only be presented here to the extent necessary to facilitate discussions of model assumptions, variable selection techniques, and results. Errors resulting from violations of model assumptions, and also from use of a binary dependent variable and intercorrelated independent variables will be presented. We will conclude with discussions of the interpretation of results for a regression model and principal component analysis.

a. Linear models

Since the exact form of relationships between dependent and independent variables is unknown, a common assumption (and good starting point) is that of a linear relationship of the form

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_m x_{im} + \varepsilon_i. \qquad (1)$$

In this study, Y_i, the dependent variable, indicates a yes-no occurrence of thunderstorms for a given time interval during a day and given combination of grid points by assuming values of one and zero, respectively. The independent variables, x's, are obtained from the measured or analyzed observations. The error or residual term, ε_i, is due to the fact that the occurrence of thunderstorms cannot be precisely predicted. The β_j's are the partial regression coefficients which relate observed conditions to the occurrence of thunderstorms. These coefficients are estimated from the data so as to minimize the sums of squared

differences between actual and estimated values of the dependent variable. Estimates of the β_j's are denoted by $\hat{\beta}_j$. This latter procedure amounts to minimizing the following:

$$\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_m x_{im})^2. \qquad (2)$$

This term is called the sum of squares of the errors or SSE. Differentiating (2) with respect to $\hat{\beta}_0$, $\hat{\beta}_1$, ... $\hat{\beta}_m$ and setting each equal to zero, we get a set of Normal Equations which can be written in matrix notation

$$(X'X)\underline{\beta} = X'\underline{Y}, \qquad (3)$$

where capital letters are matrices, underlined terms are vectors and a prime denotes the transpose of a matrix. Here (X'X) is the sum of squares and cross products of all independent variables and is called the variance-covariance matrix since we are dealing with corrected (mean subtracted from each observation) values. From (3) one can see that $\underline{\beta}$ can be obtained by multiplication of X'Y by the inverse, $(X'X)^{-1}$. The partial regression coefficients, β_j's, indicate the change in Y associated with a unit change in x while all other x's remain constant. The fact that Y is a binary variable makes no difference in these calculations.

b. Partitioning sums of squares

The statistical analysis continues by partitioning sums of squares in the fashion of analysis of variance (ANOVA) to determine the significance of the analyses as a whole as well as that of individual coefficients. The total (corrected) sums of squares can be partitioned as follows:

$$\Sigma(Y-\bar{Y})^2 = \Sigma(Y - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_m x_{im})^2 + \Sigma(\hat{\beta}_0 + \hat{\beta}_i x_{i1} + \ldots + \hat{\beta}_m x_{im} - \bar{Y})^2. \quad (4)$$

The term on the left is simply n times the variance of Y or total sum of squares (SST). The first term on the right is the sum of squared deviations of observed data from estimates based on the model. It is the residual sum of squares of the errors (SSE). The last term represents the sum of squared differences between the model estimates and estimates when no model is assumed. This is usually called the sum of squares due to regression (SSR). A mean square regression (MSR) and mean square error (MSE) are obtained by dividing SSR and SSE by their respective degrees of freedom. The partitioning is summarized in Table 1. The ratio of MSR/MSE forms the basis for the statistical F-test for

Table 1. Analysis of variance.

Source of variation	Degrees of freedom	Sum of Squares	Mean Squares	F
Total	n-1	$SST=\Sigma(Y-\bar{Y})^2$	--	--
Regression	m	$SSR=\underline{\beta}X'\underline{Y}$	SSR/m	MSR/MSE
Residual	n-m-1	SSE=SST-SSR	SSE/n-m-1	--

an hypothesis that there is no linear relation or that $\underline{\beta}=0$. Another ratio used in regression analysis is the ratio of the sum of squares of regression to the total sum of squares, SSR/SST. This quantity is sometimes called the coefficient of determination and its symbol is R^2. We can interpret R^2 as the fractional amount of total variance

accounted for by the linear combination of variables. The significance of individual partial $\underline{\beta}$'s can also be examined, but can be misleading when x's are interrelated; that is, when it is impossible to vary one x and hold all others constant. This problem will be examined in paragraphs d and e.

c. Model assumptions and violations

1) A linear model correctly describes the data.

The correct model is not known. Even if the model is of the form in (1), which parameters should be included? Variable selection techniques aid in this choice but do not guarantee that the best[2] subset has been chosen.

Within the framework of linear regression non-linear predictors are included. Linear regression refers to linear parameters ($\underline{\beta}$'s), not linear independent variables. It is unlikely that all predictors are exactly linearly related to the occurrence of thunderstorms. Fortunately, in a rather broad range for many predictors, the linear approximation is representative of the association between dependent and independent variables. We can linearize them, if we choose, by replacing the original variable by a transformed version more nearly linearly related to the predictand; however, we are not sure about its behavior when it coexists in the model with other predictors. In several attempts to linearize predictors, the overall improvement in R^2 was less than 3.0%. Also, once predictors are linearized, the equations are more

---

[2]Best or optimum refers to the maximum possible reduction of variance that can be achieved with the given linear combination of variables.

difficult to use in an operational environment. Finally, other errors to be discussed next appear to be more serious. Therefore, linearization was not pursued in this research.

2) The x's are measured without error.

We know that there are errors in measuring all variables. Not only are there errors in measuring basic variables such as temperature and wind, but also there are errors due to finite difference approximations. Unfortunately, the original data spacing and analysis procedure limit the smallest space interval for which unique information is available. Measurement error is not a problem in this study because it is small compared to the total variance of the x's. For example, the temperature error may be 0.5 K whereas the range of temperature may span 50 K.

3) The values of $\underline{\varepsilon}$ are independent, random, normally-distributed variables with a mean of zero and constant variance.

This term is estimated by residuals or differences between observed and predicted values from the computed linear function. Each item will be discussed separately.

Independent $\underline{\varepsilon}$: Meteorological variables are functions of time; however, the time dependency in our case is somewhat masked because we input data from a sequence of 36, 30, 30, 36, ... grid points for successive days. In other words, day 1 contains 36 data points; days 2 and 3 contain 30 points; day 4 contains 36 points, etc. This scheme is a consequence of the data input algorithm and remained the same for all days in this study. Also, which time dependence (one day, two days, etc.) is important? This dependence probably changes with different synoptic situations, and the overall effect is masked by other problems

to be discussed.

Randomly distributed $\underline{\varepsilon}$: A correctly specified model should show residuals which are random when plotted against an independent variable. While plots show definite non-randomness, it is most likely due to many peculiarities resulting from a dichotomous, dependent variable. These will be discussed in paragraph 5. Violations of the assumption that $\underline{\varepsilon}$ is randomly distributed are called specification error and result from not knowing the correct model form and not including the correct variables. The residual sums of squares, SSE, is, therefore, inflated and estimates of regression coefficients may be biased. There is no good way of dealing with this problem except to recognize possible nonlinearities in predictors and include physically relevant parameters. We are naturally constrained in this latter work by our fixed observational networks.

$\underline{\varepsilon}$ of constant variance, mean of zero: It is assumed that $\underline{\varepsilon}$'s are from a single population with zero mean and variance σ^2. The mean is zero but variance is a function of the x's due to the nature of the predictand in the sample used in this study. This error is termed heteroskedasticity.

Next, we will consider these last few problems in more detail.

5) Special problems for a dependent, binary variable.

In addition to the error of specification, there are several problems unique to the use of a binary dependent variable. The first and most obvious is that the error term can assume just two values depending on whether the predicted value is subtracted from zero or one. Fig. 1 is a plot of residuals for a typical predictor, W, which is positively

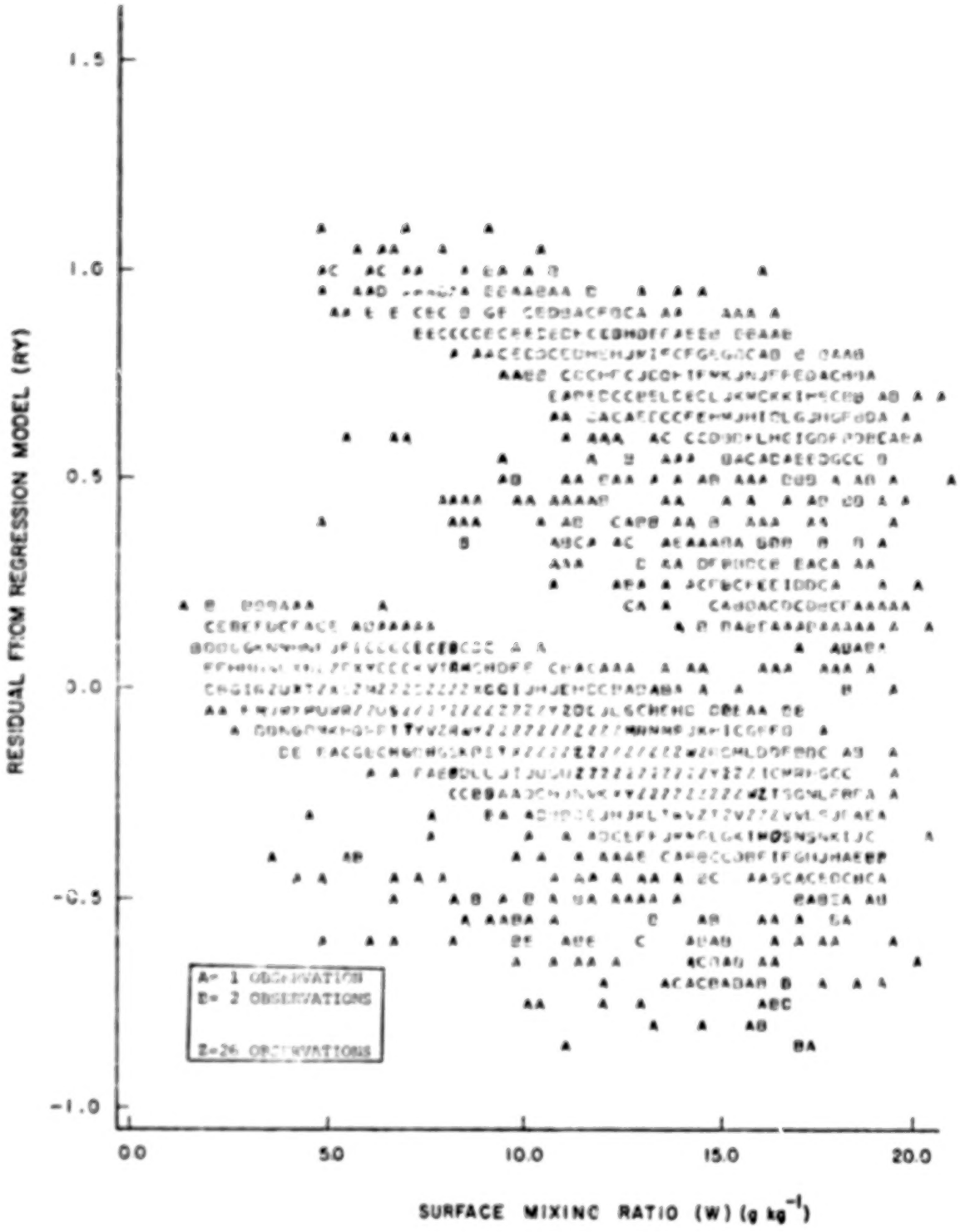


Fig. 1. Plot of mixing ratio versus residual from regression model.

correlated to thunderstorm occurrence. Each letter represents the number of observations corresponding to its position in the alphabet. A is one observation; Z represents 26 observations. Errors are clustered around small negative values (when Y is zero) and medium positive values (when Y is one and the predicted values are weighted toward zero due to the influence of all the zero observations). Obviously, the assumption of normality is not valid. The second problem is that the variance of ε_i is a function of $\mathbf{x}_i$ (Neter and Wasserman, 1974). Finally, since Y_i is similar to a probability of occurrence[3], this number should lie between zero and one. The regression response function does not automatically possess this property. Figure 2 is a plot of residuals versus predicted values for the dependent sample. Predicted values range from -0.2 to 1.2, but the mean is about 0.15.

Concerning the first problem, even though error terms are not normal, the least squares procedure still provides unbiased estimates. Further, when sample sizes are large, the distribution of estimates is asymptotically normal so that inferences concerning the regression coefficients and mean responses can still be made. Variable selection procedures, then, can still produce satisfactory results though little mention of "significance" will be made in this work. The second problem can be dealt with through weighted regression (Neter and Wasserman, 1974). Weights are assigned to observations in such a way that responses or predicted values near zero or one receive maximum weight.

---

[3]We are trying to predict an occurrence which is represented by a one or a non-occurrence represented by zero in a continuous fashion.

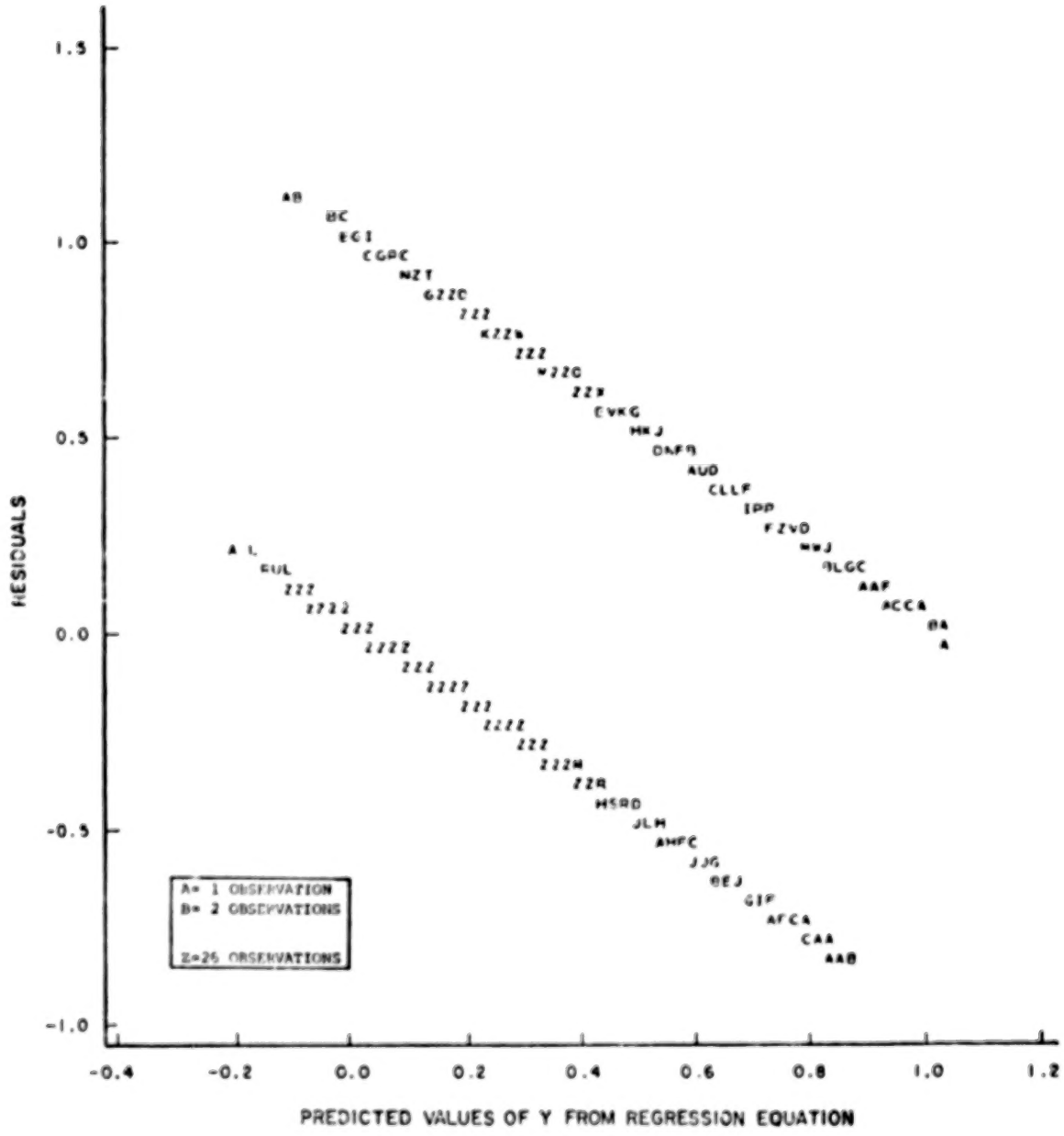


Fig. 2. Plot of residuals versus predicted values from regression model.

This type of regression was not performed because the observations of thunderstorms are already weighted by virtue of low climatological frequencies of thunderstorm occurrence. An attempt was made to deal with the problem through inclusion of random samples of no-thunderstorm observations and through prior screening of no-thunderstorm cases by critical values of selected predictors. The problem of predicting less than zero or greater than one is not particularly serious since the threshold for forecasting thunderstorms from predicted values is arbitrary. Nevertheless, it appears that fitting a logistic function such as

$$Y = (\exp(-10.0 + 0.1x))/(1 + \exp(-10.0 + 0.1x)) \qquad (5)$$

would eliminate this problem. Such a function is shown in Fig. 3 for one independent variable.

It can be linearized by the simple transformation,

$$Y' = \ln (Y/(1-Y)). \qquad (6)$$

Special precautions are required for zero predicted values. Note that here, too, added weight is given to both near-zero and one predictors. Glahn and Bocchieri (1975) used a similar function in an objective forecasting scheme and found difficulties in some cases due to the symmetric nature of the curve and poor fit near the threshold probability for yes-no forecasts. Also, fitting this function is not easy unless there are repeat observations for each level of x. Such is not the case with the data used in this research.

d. Multicollinearity

Another more serious problem results from use of interrelated predictors (x's). The x's are in fact related in at least three ways.

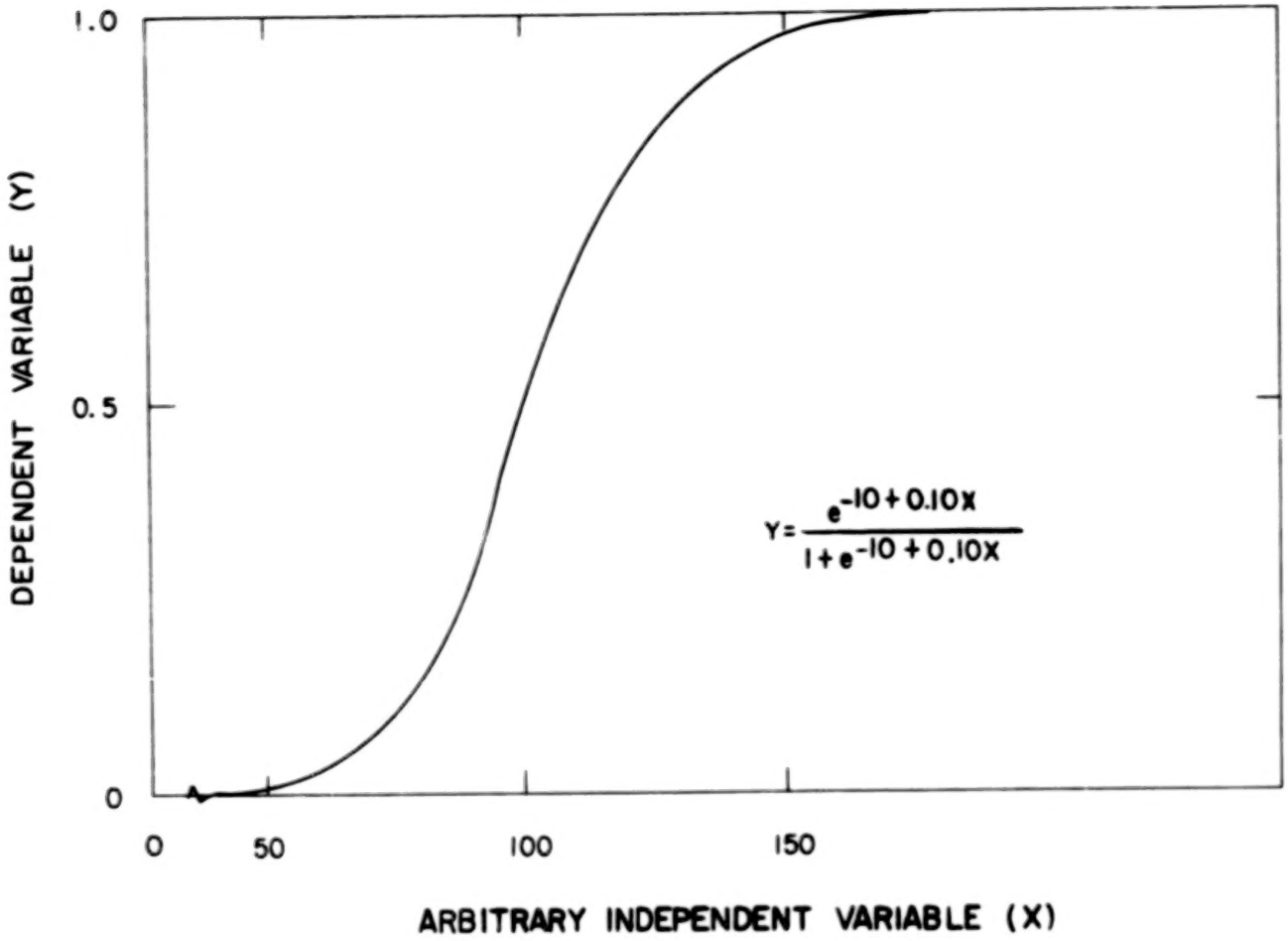


Fig. 3. A logistic response function.

First, the basic, measured variables are related through physical laws and relationships such as the gas law, first law of thermodynamics, or thermal wind equation; therefore, parameters derived from the basic variables are related. This problem is usually exaggerated by using the basic five surface variables and five upper-air variables (the latter for each of four chosen pressure levels in the troposphere), and computing up to 55 parameters for more than twice as many points in space as there are original data measurements[4]. Also, several measures of the same basic dimension, say stability, are calculated because the best measure of stability is not known. Therefore, many more variables than we need are included. Secondly, variables are related in space for many hundreds of kilometers. The very concept of an air mass suggests a dependence for many variables. Finally, there is a time dependence in that meteorological variables on one day are correlated to those on the next day (or longer).

The problem of intercorrelated "independent" variables is called multicollinearity and for data in this research is severe enough to prevent us from calculating $(X'X)^{-1}$ since near-singularities exist[5]. Therefore, we must use a variable-selection technique to be discussed in Section 2e or a principal component analysis discussed in Section 2g. When an inverse can be computed and the model is correct, then the regression coefficients estimated by the least squares technique are

---

[4]This is a consequence of the analysis schemes, and the price paid for trying to preserve as much detail as possible.

[5]A generalized inverse can be calculated; however, the estimates of the coefficients would be biased.

unbiased. This means that the expected values computed from repeated samples will approach the correct value in the mean. The space-correlation problem is reduced by use of every fourth grid point in the statistical analyses.

e. Variable selection methods and inference

Fortunately, the regression analysis is robust in that even moderate deviations from the assumptions do not invalidate results. The significant problem of multicollinearity, however, can have severe influences (even critical in our case with all variables where the X'X matrix is near singular). Of many interrelated variables, which should be kept in the model? To deal with this problem, four different variable selection techniques were used in this study; all try to choose subsets of predictors which minimize the residual mean square (MSE). They are forward selection, backward elimination, stepwise, and maximum R^2 improvement.

1) Forward selection

This procedure, often called step-up, begins by choosing that variable which is most highly correlated with the dependent variable. The second variable is chosen by seeking the next most highly correlated of the remaining independent variables with the dependent variable, according to the partial correlation coefficient. In other words, for each remaining independent variable, a partial F-statistic is calculated that reflects that variable's contribution to the model were it to be included. If this statistic for one or more variables has a "significance level" greater than a specified amount (0.50 is used in this study), then the variable with the largest F is included. This process

is repeated and variables added one at a time until none passes the F-test or no more remain. Once a variable is added to the model, it must remain whether or not its influence is negated by other variables added. This procedure is likely to give near optimum few-variable models, but deteriorates as more are added.

2) Backward elimination

In this technique, also called step-down, the model with all variables is considered; then variables are deleted one at a time starting with the one whose β exhibits the lowest F-statistic. Here, we are likely to get optimum many-variable models but poor results when more and more variables are deleted since they can never be included again.

3) Stepwise

This procedure is a refinement of forward selection. At each step before determining the next variable to be added, the F-statistics are checked for the coefficients already chosen to see if any should be deleted based on another prespecified "significance level" (in our case 0.1). Only after this check for deletion is made can another variable be added. The procedure terminates when no partial F is ≥ 0.5 or a variable to be added is one just deleted. This procedure is most appealing so far; but, still an optimum subset is not guaranteed (Draper and Smith, 1966). Stepwise is the predominant procedure used in this research.

4) Maximum R^2 improvement

A one-variable model is chosen as with forward selection. Then every combination of variables with this one is examined. When two variables are included each of these is compared to each variable not

in the model. For every comparison it is determined if removing the variable in the model and replacing it with the excluded variable would increase R^2. After all comparisons, the switch is made that gives the highest R^2. This process continues with each variable added. Optimum one-to-eight variable models are most likely to be found, but the costs in computer processing are high when more than 20 candidate predictors are used (Barr et al., 1976).

Although variable selection procedures do not guarantee that an optimum subset of predictors is chosen, the stepwise procedure does a credible job up to about the fourth variable for data in this study. Comparisons were made of variables selected by the stepwise procedure with those from the best four- to seven-variable models where all possible regressions were considered. In all cases the four-variable models were identical. The five- and six-variable models differed by just one variable. The best seven-variable model differed by two variables. Due to computer-processing limitations, comparisons were not exact in that only 18 of 25 predictors were considered for all possible regressions. Even for this combination there were 31,824 possibilities. In the case of the seven-variable model, the two variables not selected by the "best" procedure were not available to it. Beyond five predictors there could be any number of variable combinations which produce the same or even slightly higher R^2. Therefore, discussions of variable combinations will usually be limited to the first four or five.

f. Interpretation of regression-model results

In these discussions, intense convection, thunderstorm occurrence, and MDR≥4 are used synonymously, though the latter is the true

predictand. The coefficient of determination, R^2, the amount of variance accounted for by the linear combination of variables, and reduction of total variance due to the regression model also are used synonymously. Finally, independent variables, predictors, and x's mean the same as do dependent variable, Y, and predictand.

Models of form (1) are used where particular x's, parameters, are chosen by variable selection techniques discussed in Section 2e. The associated coefficients, β_j's, are computed according to the least squares method (Section 2a). Analysis of variance tables such as shown in Table 1 (p.12) are produced for every different combination of independent variables and for all data subdivisions. A few of these tables for important combinations of parameters are shown in Appendix A. In general, however, only summaries are included in the text. These summaries present the total R^2, number of variables (x's) which produced the R^2, mean square error for this number of variables, and occurrence frequency for the dependent variable (frequency of thunderstorm occurrence). Also shown are the variables selected in the order in which they were chosen, the cumulative R^2, and the sign of the partial regression coefficient (β) for each data stratification. In order to reconstruct the linear equation for a given combination of variables, the partial regression coefficients from Appendix A are required. These coefficients are then substituted into (1) together with their respective predictors.

Although R^2 will be discussed to some extent, the R^2 differences from sample to sample must not be interpreted to imply improved regression results unless the proportion of ones (as opposed to zeros) is

also the same. For a binomial distribution the variance is given by np(1-p). Since this term appears in the denominator of R^2, an increased p (up to 0.5) results in a lower R^2 given the same regression sum of squares. Three examples follow, each using similar but artificial data with one independent, continuous variable positively correlated to one dependent, dichotomous variable.

1) Example one: occurrence frequency 10%

Assume there are ten observations of dependent variable y and independent variable x and that the frequency of ones is 10%. The data and regression analysis are shown in Table 2.

Table 2. Data and regression analysis for 10 observations of hypothetical variables x and y with 10% occurrence frequency.

y	$(y-\bar{y})$	$(y-\bar{y})^2$	x	$(x-\bar{x})$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$	$[(x-\bar{x})(y-\bar{y})]^2$	Summary Statistics
0	-0.1	0.01	2	-1	1	0.1	0.01	
0	-0.1	0.01	2	-1	1	0.1	0.01	$SSR = \frac{\Sigma[(x-\bar{x})(y-\bar{y})]^2}{\Sigma(x-\bar{x})^2} = 0.330$
0	-0.1	0.01	2	-1	1	0.1	0.01	
0	-0.1	0.01	3	0	0	0.0	0.00	$SST = \Sigma(y-\bar{y})^2 = 0.9$
0	-0.1	0.01	3	0	0	0.0	0.00	
0	-0.1	0.01	4	1	1	-0.1	0.01	$R^2 = \frac{SSR}{SST} = .367$
0	-0.1	0.01	2	-1	1	0.1	0.01	
1	0.9	0.81	5	2	4	1.8	3.24	$\bar{x} = 3.0, \quad \bar{y} = 0.1 = p$
0	-0.1	0.01	3	0	0	0.0	0.00	
0	-0.1	0.01	4	1	1	-0.1	0.01	
1		0.90	30		10		3.30	Sum

2) Example two: occurrence frequency 30%

Table 3 illustrates another example with the same sample size but different occurrence frequency. Note that as the occurrence frequency increases, R^2 decreases. This decrease is a consequence of the increased variance of y (higher p).

Table 3. Data and regression analysis for 10 observations of hypothetical variables x and y with 30% occurrence frequency.

y	$y-\bar{y}$	$(y-\bar{y})^2$	x	$x-\bar{x}$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$	$[(x-\bar{x})(y-\bar{y})]^2$		Summary Statistics
0	-0.3	0.09	2	-1	1	0.3	0.09		
1	0.7	0.49	4	1	1	0.7	0.49		$SSR = \frac{\Sigma[(x-\bar{x})(y-\bar{y})]^2}{\Sigma(x-\bar{x})^2} = 0.29$
1	0.7	0.49	4	1	1	0.7	0.49		
0	-0.3	0.09	2	-1	1	0.3	0.09		$SST = \Sigma(y-\bar{y})^2 = 2.10$
0	-0.3	0.09	3	0	0	0	0		
0	-0.3	0.09	4	1	1	-0.3	0.09		$R^2 = \frac{SSR}{SST} = 0.138$
0	-0.3	0.09	2	-1	1	0.3	0.09		
1	0.7	0.49	5	2	4	1.4	1.96		$\bar{x} = 3.0; \bar{y} = 0.3 = p$
0	-0.3	0.09	2	-1	1	0.3	0.09		
0	-0.3	0.09	2	-1	1	0.3	0.09		
3		2.10	30		12		3.48	Sum	

3) Example three: random sampling

We will now consider the effect of random sampling on R^2 in Table 4. Table 3 is duplicated for all occurrences but for only 57% of the nonoccurrences.

Table 4. Data and regression analysis for 57% of nonoccurrence observations in Table 3 data.

y	$(y-\bar{y})$	$(y-\bar{y})^2$	x	$x-\bar{x}$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$	$[(x-\bar{x})(y-\bar{y})]^2$		Summary Statistics
0	-0.429	0.184	2	-1.143	1.306	0.490	0.240		
1	0.571	0.329	4	0.857	0.735	0.489	0.239		$SSR = \frac{\Sigma[(x-\bar{x})(y-\bar{y})]^2}{\Sigma(x-\bar{x})^2} = .263$
1	0.571	0.326	4	0.857	0.735	0.489	0.239		
0	-0.429	0.184	3	0.143	0.020	-0.061	0.004		$SST = \Sigma(y-\bar{y})^2 = 1.714$
0	-0.429	0.184	2	-1.143	1.306	0.490	0.240		
1	0.571	0.326	5	1.857	3.449	1.060	1.124		$R^2 = \frac{SSR}{SST} = 0.154$
0	-0.429	0.184	2	-1.143	1.306	0.490	0.240		
									$\bar{x} = 3.14; \bar{y} = 0.429 = p$
3		1.714	22		8.857		2.328	Sum	

Table 5 shows the comparison of pertinent statistics for the different occurrence frequencies and the random sample. In the case of the random sample, sums of squares of x's decrease relative to the other cases because $\bar{x}$ increases. Total sum of squares, $\Sigma(y-\bar{y})^2$, decreases compared to the 30% sample in this case because there are fewer elements to sum. Finally, the squared sum of cross products

Table 5. Comparison of Tables 2, 3, and 4.

Term	Table 2 10% ones	Table 3 30% ones	Table 4 Random (43% ones)
$\bar{x}$	3.0	3.0	3.140
$\Sigma(x-\bar{x})^2$	10.0	12.0	8.857
$\bar{y}$	0.1	0.3	0.429
$\Sigma(y-\bar{y})^2$	0.9	2.1	1.714
$\Sigma[(x-\bar{x})(y-\bar{y})]^2$	3.30	3.48	2.328
SSR	0.330	0.290	0.263
R^2	0.367	0.138	0.154
MSE	0.071	0.226	0.289
n	10	10	7

decreases but at a slower rate than in Table 3. Consequently, R^2 increases for the random sample compared to the 30% case. It is clear from these examples that R^2 cannot be used as a measure of relative strength of the regression model when the frequency of occurrence changes. The only true measure of "goodness" will be the performance of the function in an operational environment.

g. Principal component analysis

Another way to approach the multicollinearity problem is through a technique called principal component analysis first introduced to meteorology over two decades ago by Lorenz (1956). Brier and Meltesen, (1976) give a brief history of meteorological applications. Only a summary of the methodology will be presented here.

Assume that new variables, principal components (C_i), can be generated that are linear combinations of observations of original variables as follows:

$$
\begin{aligned}
C_1 &= b_{1,1}x_1 + b_{1,2}x_2 + \ldots + b_{1,m}x_m \\
C_2 &= b_{2,1}x_1 + b_{2,2}x_2 + \ldots + b_{2,m}x_m \\
&\vdots \qquad\qquad\qquad\qquad\qquad \vdots \\
C_m &= b_{m,1}x_1 + b_{2,m}x_2 + \ldots + b_{m,m}x_m
\end{aligned}
\tag{7}
$$

Also choose coefficients for C_i i.e. b_{ij}'s so that the variance of C_1 is as large as possible. Choose the C_2 coefficients so that the variance of C_2 is as large as possible subject to the constraint that observations of C_1 be uncorrelated with those of C_2. We continue for all C_i and impose an additional restriction that squares of coefficients in any C_i sum to one. It turns out (Harris, 1975) that if the eigenvalues and eigenvectors of the X'X matrix are found (since it is real, symmetric, and positive definite), then the assumptions are fulfilled. Also, the components of the eigenvectors normalized to length one are the b_{ij}'s.

Since the variance is just a measure of the variability for different observations, it is reasonable to interpret C_1 as that linear combination of original variables which maximally discriminates among our observations. These components also partition the total variance of the original variables into m additive parts, hence, the interpretation that they "account for" a certain fraction of the total variance. Rows (or columns) in the symmetric (X'X) matrix which are linear combinations of each other will produce a zero eigenvalue and will contribute nothing to the total variance; hence, we have another way of assessing multicollinearity and of finding, possibly, how many true dimensions or hypothetical latent variables there are in the particular (X'X) matrix which is evaluated in this manner. It is this property which

has led to recent applications in meteorology (Smith and Woolf, 1976; Brier and Melteson, 1976). A method for calculating eigenvalues is given by Essenwanger (1976). The procedures used in this study are those available in the statistical analysis system (SAS) (Barr, et al., 1976).

3. DATA SELECTION AND PROCESSING

a. Location

The area for this study was chosen to provide relative homogeneity in terrain, an adequate sample of meteorological observations, and as many thunderstorm occurrences as possible during the time digital radar data were available. The period chosen included April through July 1974 and 1975, 30 days in each month. Surface, upper-air and meteorological radar data were used in the analysis. Each will be discussed separately.

b. Surface data

Altimeter setting, wind speed, wind direction, temperature, and dew point temperature were obtained for 97 locations as shown in Fig. 4 for five times each day: 1200, 1500, 1600, 1700, and 1800 GMT.

c. Upper-air data

Observations of geopotential height, temperature, dew point depression, wind speed, and wind direction at 1200 GMT were used for each of four standard pressure levels: 850, 700, 500, and 300 mb[6]. There were 14 upper-air locations (Fig. 4). Both surface and upper-air data were obtained from the USAF Environmental Technical Applications Center at Scott AFB, IL.

d. Radar data

Radar data consisted of manually digitized radar (MDR) observations

---

[6]Only geopotential height and winds were utilized for the 300-mb level.

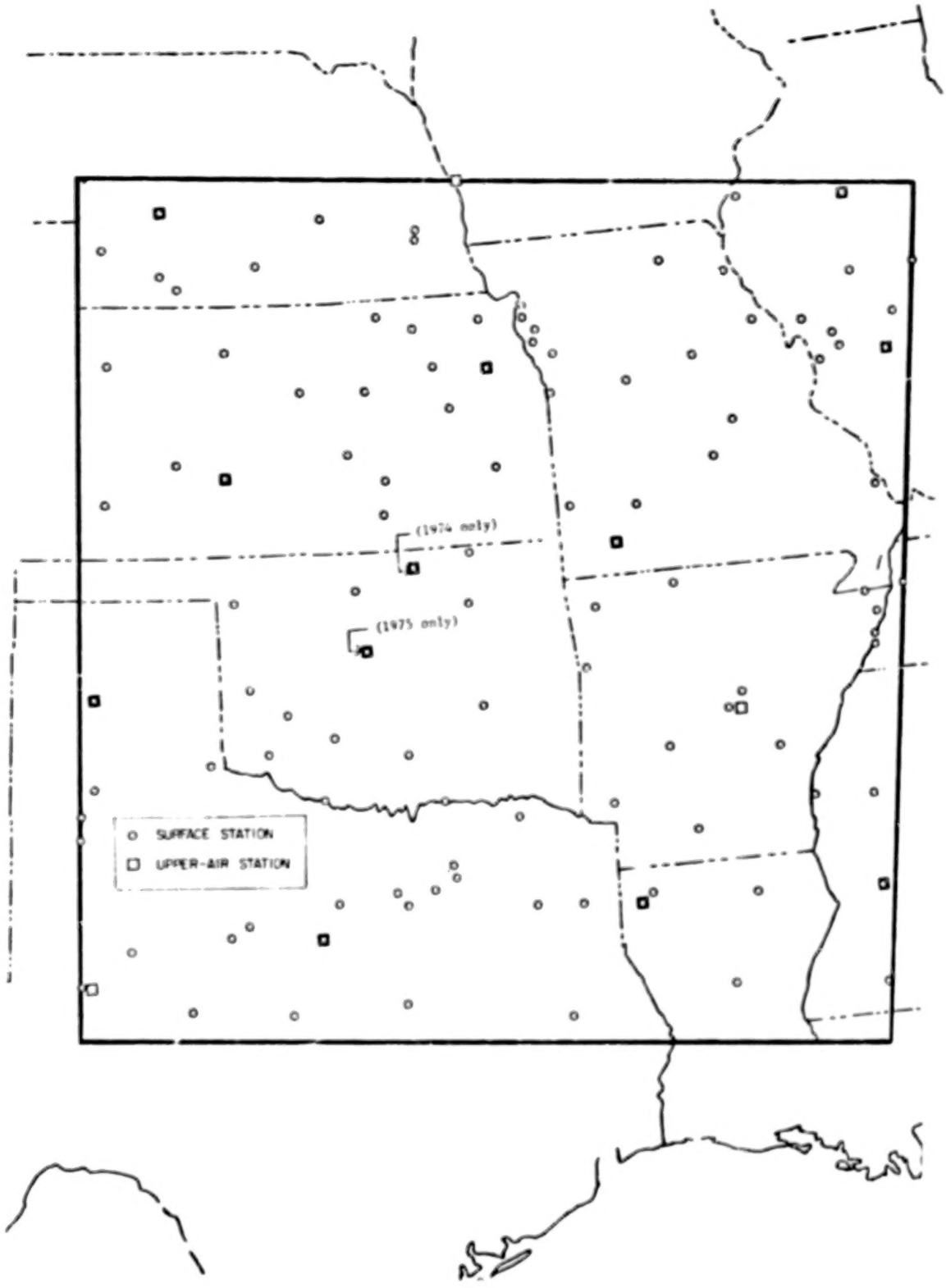


Fig. 4. Data reporting locations.

for each hour from 1630 to 0130 GMT and for 187 boxes shown within the bold line in Fig. 5. Note that the centers of these boxes fall within the general area outlined in Fig. 4 (p. 32). These data were provided by NOAA's Techniques Development Laboratory. Radar observations are usually taken about 30 to 35 min past each hour and transmitted in coded form (Table 6) (Foster and Reap, 1973). Digital codes represent the maximum

Table 6. Explanation of Manually Digitized Radar (MDR) code

Code No.	Maximum Observed VIP[1] Values	Coverage In Box	Maximum Rainfall Rate (in./hr)	Intensity Category
0	No Echoes			
1	1	Any VIP1	< 1	Weak
2	2	≤ 50% of VIP2	.1- .5	Moderate
3	2	> 50% of VIP2	.5-1.0	Moderate
4	3	≤ 50% of VIP3	1.0-2.0	Strong
5	3	> 50% of VIP3	1.0-2.0	Strong
6	4	≤ 50% of VIP3 and 4	1.0-2.0	Very Strong
7	4	> 50% of VIP3 and 4	1.0-2.0	Very Strong
8	5 or 6	≤ 50% or VIP3, 4, 5, and 6	>2.0	Intense or Extreme
9	5 or 6	> 50% or VIP3, 4, 5, and 6	>2.0	Intense or Extreme

[1]Video Integrator Processor

intensity of reflectivities anywhere in a square area approximately 85 km on a side. These codes also take into account the general area coverage of the echoes. Time composites for the maximum code in any of the following groups were saved for each day: 1635-1735, 1835-1935, 1935-2235, and 2235-0135 GMT. These will be called 1700-1800 GMT,

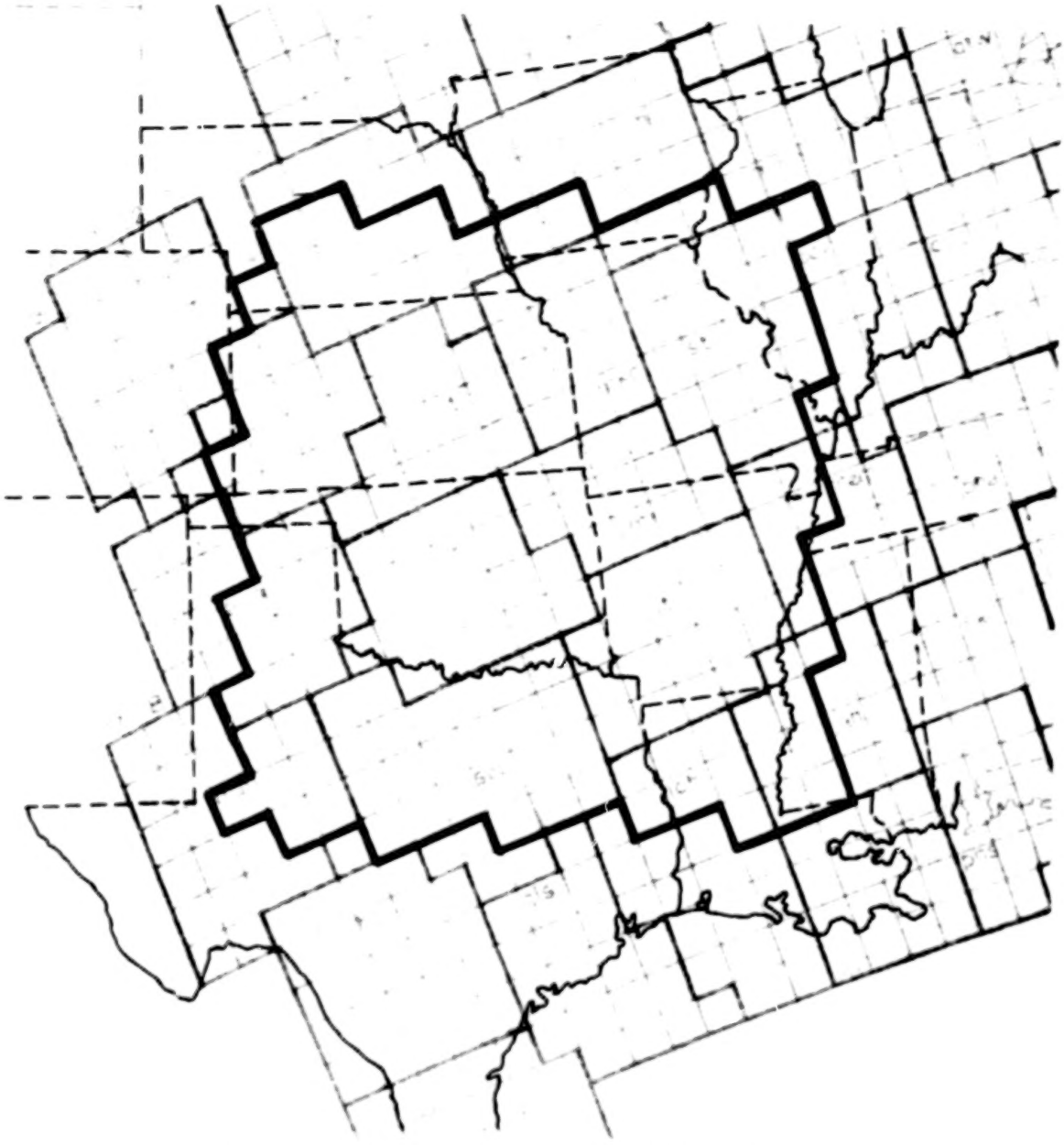


Fig. 5. Manually digitized radar grid.

1900-2000 GMT, 2000-2300 GMT and 2300-0200 GMT periods, respectively. Radar data had to be grouped by time intervals to obtain an adequate sample because many hours of observations were missing. There are several reasons for the specific groupings. The first interval is to be used as a candidate predictor. The latter three were all predictands and were formulated from operational considerations. A 3-h interval represents a forecast of thunderstorms valid within 1.5 h of an estimated time of arrival for aircraft flight operations. For example, an aircrew may obtain a weather briefing at 1830 GMT for a 5.5-h flight with departure time estimated to be 1900 GMT and estimated arrival time at destination of 0030 GMT. A forecast for intermittent thunderstorms would cover the period from 2300 to 0200 GMT. This interval is reasonable owing to operational uncertainties such as delays in departure and landing for long flights and to uncertainties in predicting the thunderstorm event so long in advance. The 1-h interval at the earlier time reflects both reduced forecast and operational uncertainties because of the short forecast lead time and brief flying time. For example, a crew for a 1-h flight may get a weather briefing at 1800 GMT for estimated arrival at 1930 GMT. The forecast would then cover the period from 1900 to 2000 GMT. Finally, an attempt was made to avoid overlapping intervals so that forecasts for the different times could be compared.

e. Initial processing

Raw data were available on magnetic tapes. Programs were written to (1) select specific observed elements, times, and stations; (2) ensure all missing hours and days were accounted for; (3) check for gross errors in reported values; and (4) write all data onto a direct access

storage device. Observations that were either missing or which contained numbers outside the range of what would be considered reportable values for that variable were filled with zeros and ignored in subsequent processing. Many observations were checked against archived teletype data to ensure accuracy.

f. Objective analysis

The results of this research were dependent upon the representativeness of raw data interpolated or analyzed onto an equally-spaced grid system. Therefore, considerable care was taken in choosing an analysis procedure and grid. An 18 x 18 array of grid points spaced 65 km apart was chosen to preserve as much detail in the surface and radar data fields as possible. Boundary points were used only for the calculation of derivatives so that only 256 points (16 x 16) were used for statistical correlations. An analysis scheme by Barnes (1964) was selected, not only because results obtained were very similar to hand analysis, but also because scales of atmospheric features retained by this technique could be determined, and the program was efficient. Scan radii and initialization procedures were adjusted to produce an optimum balance among the following: (1) cost, since we had 12,480 total analyses to perform[7]; (2) missing data; (3) amplification of spurious waves; (4) small-scale surface features; (5) radar grid transposition; and (6) duplication of manual analyses. The optimum choice for scan

[7] 240 days x (5 surface variables x 5 times + 5 upper-air variables x 3 levels + 3 upper-air variables x 1 level + 1 radar variable x 9 times).

radius, number of iterations, and characteristic wave lengths preserved as a result of these choices are summarized in Table 7. Wind was converted to components with respect to grid orientation (nearly latitude-longitude aligned). These and all other basic variables were analyzed onto the 18 x 18 grid array for each time and day. From these data the predictand and candidate predictors were computed at each grid point as discussed in the next Section.

Table 7. Summary of analysis parameters.

Data Source	Average data Spacing	Scan Radius	Iterations	Initialization	Wavelength of 90% amplitude Preservation	Wavelength of 50% amplitude Preservation
Surface	120 km	275 km	3	Mean value of parameter	450 km	300 km
Upper air	370 km	520 km	3	Mean value of parameter	900 km	600 km
Radar	83 km	84 km	1	0	*	*

*With one iteration this was essentially an interpolation of the nearest MDR observation to each grid point.

4. PARAMETERIZATION AND DATA SUBDIVISION

This section includes the formulation of predictands and the development of predictors in the context of parameterization of synoptic observations. Also discussed is the subdivision of the total data set into subsets for statistical processing.

a. Predictand formulation

Coded MDR data from the 65-km grid and three time groups, 1900-2000 GMT, 2000-2300 GMT, and 2300-0200 GMT, were converted to a simple binary form. Any MDR code equal to or greater than four at any of the four nearest neighbor grid points as shown in Fig. 6 was assumed to represent the occurrence of a thunderstorm (Mogil, 1974), and was assigned the binary code one; otherwise code zero was assigned. The data void areas in this figure result from the use of every fourth grid point for the statistical analyses. A zero could only be assigned if the grid point in question and the nearest neighbors were all reporting MDR codes less than four. The best resolution in the predictand area is limited to a square area about 138 km on a side. This was the smallest area for which unique information from the original radar grid (83 km square) was available. We have not distinguished among precipitation intensities (or thunderstorm severities) in this study.

b. Predictor formulation

One approach now tempting many investigators because of expanded computer capabilities is to use every imaginable parameter as a candidate predictor. For just the basic analyzed variables (temperature, wind components, pressure, etc.) along with their first and second time

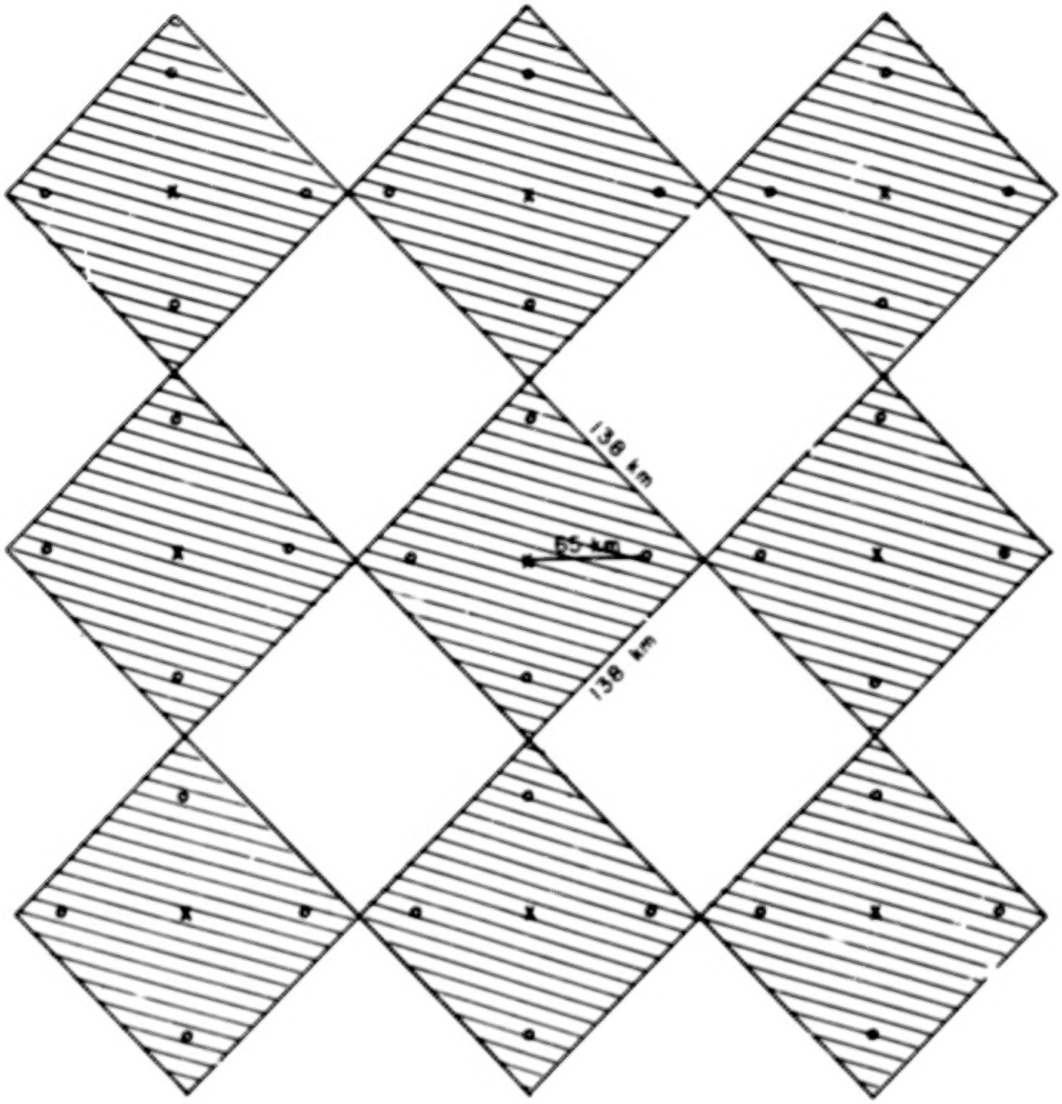


Fig. 6. Orientation of the predictand area with respect to the predictor point.

and space derivatives, there would be well over 100 candidates, many of which would be interrelated. Selection techniques for such a large number, not even counting products or time changes of space derivatives and vice versa, would be expensive and, more important, results would be extremely difficult to interpret. In this study all predictors have been chosen through parameterization techniques for categories of variables known to be associated with thunderstorms.

It is generally recognized that there are three synoptic-scale[8] conditions for thunderstorms: moisture, potential instability, and a trigger mechanism. Therefore, parameters to represent these ingredients were calculated from centered finite differences for which the distance interval was twice the grid distance or 130 km. In addition, a nine-point Laplacian routine for $\nabla^2 A$ was used, where A is any scalar. All parameters, along with their definition and source, are shown in Table 8. Each group will be discussed separately.

1) Moisture

The first set of moisture variables includes the equivalent potential temperature (θ_e) at several levels in the atmosphere, its time change, gradient magnitude, and advection. This parameter has been used for many years as a means of identifying air samples owing to its conservative properties for both dry and saturated adiabatic processes. It has been used recently in conjunction with the location of the thunderstorm updraft (see, for example, Ellrod and Marwitz, 1976; Fankhauser, 1974; Brandes, 1977). High values of θ_e represent a

---

[8] The data network restricts the horizontal scales to 300 to 1500 km.

Table 8. Candidate predictors

(a) Moisture Parameters

Symbol	Definition	Source	Time
θ_e	$\theta[\exp(LW_s/C_pT)]$	Surface	1800
θ_{e_8}	same except 850 mb values	Upper air	1200
θ_{e_7}	same except 700 mb values	Upper air	1200
$\frac{\partial\theta_e}{\partial t}$	θ_e(1800 GMT) - θ_e(1500 GMT)	Surface	[1500 / 1800]
$\frac{\partial^2\theta_e}{\partial t^2}$	$[\frac{\partial\theta_e}{\partial t}$ - (θ_e(1500 GMT) - θ_e(1200 GMT))]	Surface	[1200 / 1500 / 1800]
$\lvert\vec{\nabla}\theta_e\rvert$	$\sqrt{(\frac{\partial\theta_e}{\partial x})^2 + (\frac{\partial\theta_e}{\partial y})^2}$	Surface	1800
$\theta_e A$	$-(u\frac{\partial\theta_e}{\partial x} + v\frac{\partial\theta_e}{\partial y})$	Surface	1800
$T-T_d$	$T-T_d$	Surface	1800
$(T-T_d)_8$	same except 850 mb values	Upper air	1200
$(T-T_d)_7$	same except 500 mb values	Upper air	1200
W	$0.622e/P-e$ $e = (6.11)10^{7.5(T_d-273.18)/T_d-35.86}$ $P = -1013.25 + 1013.25/(1.0-a(z))^b$ +ALTSTG, where $a(z) = .0065z/288.0$, $b = 5.246$	Surface	1800
W_8	same except for 850 mb	Upper air	1200
W_7	same except for 700 mb	Upper air	1200
$\lvert\vec{\nabla}W\rvert$	$\sqrt{(\frac{\partial W}{\partial x})^2 + (\frac{\partial W}{\partial y})^2}$	Surface	1800
$\vec{\nabla}^2W$	$\frac{\partial^2W}{\partial x^2} + \frac{\partial^2W}{\partial y^2}$	Surface	1800
MDIV	$\vec{\nabla}\cdot W\vec{V}$	Surface	1800
$\lvert D\rvert$	$\sqrt{\frac{\partial W}{\partial x}\frac{\partial u}{\partial y} + \frac{\partial W}{\partial y}\frac{\partial v}{\partial y})^2 + (\frac{\partial W}{\partial x}\frac{\partial u}{\partial x} + \frac{\partial W}{\partial y}\frac{\partial v}{\partial x})^2}$	Surface	1800

Table 8. Candidate predictors (Continued)

(b) Stability

Symbol	Definition	Source	Time
DTA	$(-\vec{V}\cdot\vec{\nabla}_p T)_8 - (\vec{V}\cdot\vec{\nabla}_p T)_5$	Upper air	1200
CSIL	$\theta_{e_7} - \theta_{e_8}$	Upper air	1200
CSIM	$\theta_{e_5} - \theta_{e_8}$	Upper air	1200
KI	$T_8 + T_{d_8} - (T-T_d)_7 - T_5$	Upper air	1200
TTI	$2(T_8 - T_5) - (T - T_d)_8$	Upper air	1200
STSI	$\frac{RT_7}{\theta_{e_7} P_7} (\theta_{e_5} - \theta_{e_8})$	Upper air	1200
UWSH	$u_5 - u_8$	Upper air	1200
DTH	$(Z_8-Z_7)/150 - (Z_7-Z_5)/200$	Upper air	1200
$\vec{\nabla}^2_p$THA	$\vec{\nabla}^2_p(-\vec{V}_7\cdot\vec{\nabla}(\text{DTH}))$	Upper air	1200
T	Temperature	Surface	1800
θ	$T(1000/P)^{R/C_p}$	Surface	1800

Table 8. Candidate predictors (Concluded)

(c) Trigger

Symbol	Definition	Source	Time
ω_{TS}	$(-\vec{V}\cdot\vec{\nabla}z)g\rho + (\vec{\nabla}\cdot\vec{V})50$	Surface	1800
$\vec{\nabla}^2 P$	$\frac{\partial^2 P}{\partial x^2} + \frac{\partial^2 P}{\partial y^2}$	Surface	1800
ζ	$\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}$	Surface	1800
DVA	$(-\vec{V}\cdot\vec{\nabla}_p\zeta)_5 - (-\vec{V}\cdot\vec{\nabla}_p\zeta)_8$	Upper air	1200
IDIV	$2.25(\vec{\nabla}_p\cdot\vec{V})_8 + 1.75(\vec{\nabla}_p\cdot\vec{V})_7 + 1.0(\vec{\nabla}_p\cdot\vec{V})_5$	Upper air	1200
IMDIV	$(\vec{\nabla}\cdot w\vec{V})_8 + (\vec{\nabla}_p\cdot w\vec{V})_7 + (\vec{\nabla}_p\cdot w\vec{V})_5 + (\vec{\nabla}_p\cdot w\vec{V})_3$	Upper air	1200
$\lvert\vec{V}_5\rvert$	$\sqrt{u_5^2 + v_5^2}$	Upper air	1200
v_5	500 mb N-S wind component	Upper air	1200
VSUM	$v_5 + v_8$	Upper air	1200
$\lvert\vec{\nabla}P\rvert$	$\sqrt{(\frac{\partial P}{\partial x})^2 + (\frac{\partial P}{\partial y})^2}$	Surface	1800
$\vec{V}\cdot\vec{\nabla}P$	$u\frac{\partial P}{\partial x} + v\frac{\partial P}{\partial y}$	Surface	1800
$\frac{\partial(\vec{\nabla}^2 P)}{\partial t}$	$\vec{\nabla}^2 P$(1800 GMT) - $\vec{\nabla}^2 P$(1500 GMT)	Surface	$\{^{1500}_{1800}\}$
MDRP	MDR code = 1 at 1700 or 1800 GMT	Radar	$\{^{1700}_{1800}\}$

potential, latent energy source (warm moist air) for the convective process.

Next, basic measures of low-level relative humidity are included. These are expressed as dew point depressions. The last group of moisture parameters are basic measures of atmospheric water vapor content. The mixing ratio has been combined with the divergence field of surface wind in $\vec{\nabla}\cdot w\vec{V}$ so that moisture advection and convergence are included in a single term. This has been a leading predictor in other studies (Charba, 1977; Alaka et al., 1973; Henz, 1974). The Laplacian of mixing ratio identifies centers of high moisture (negative Laplacian). A term which combines both the deformation field of the wind and surface moisture pattern has been introduced in $|D|$. This is similar in form to the frontogenetic function of Petterssen (1956, p. 201) with θ replaced by W and is discussed in Palmén and Newton (1969, p. 246). It is a way of locating where shear and confluence of surface wind could concentrate moisture. Fig. 7 shows schematically how this might be accomplished. Prime quantities represent isolines after a time increment Δt. The lines of W' in Fig. 7a have been shifted to the left for convenience. Consider the magnitude of $\vec{\nabla}w$. As $\frac{\partial u}{\partial y}$ decreases, $|\vec{\nabla}w|$ increases. Similarly, as $\frac{\partial v}{\partial y}$ decreases, $|\vec{\nabla}w|$ increases.

2) Stability

There are numerous ways of estimating atmospheric static stability. Differential temperature advection where cold air is advected over warm air or vice versa is a way of incorporating kinematics (wind structure) and time. So long as the advection is constant with cold advection above warm advection, the atmosphere will respond by decreasing stability

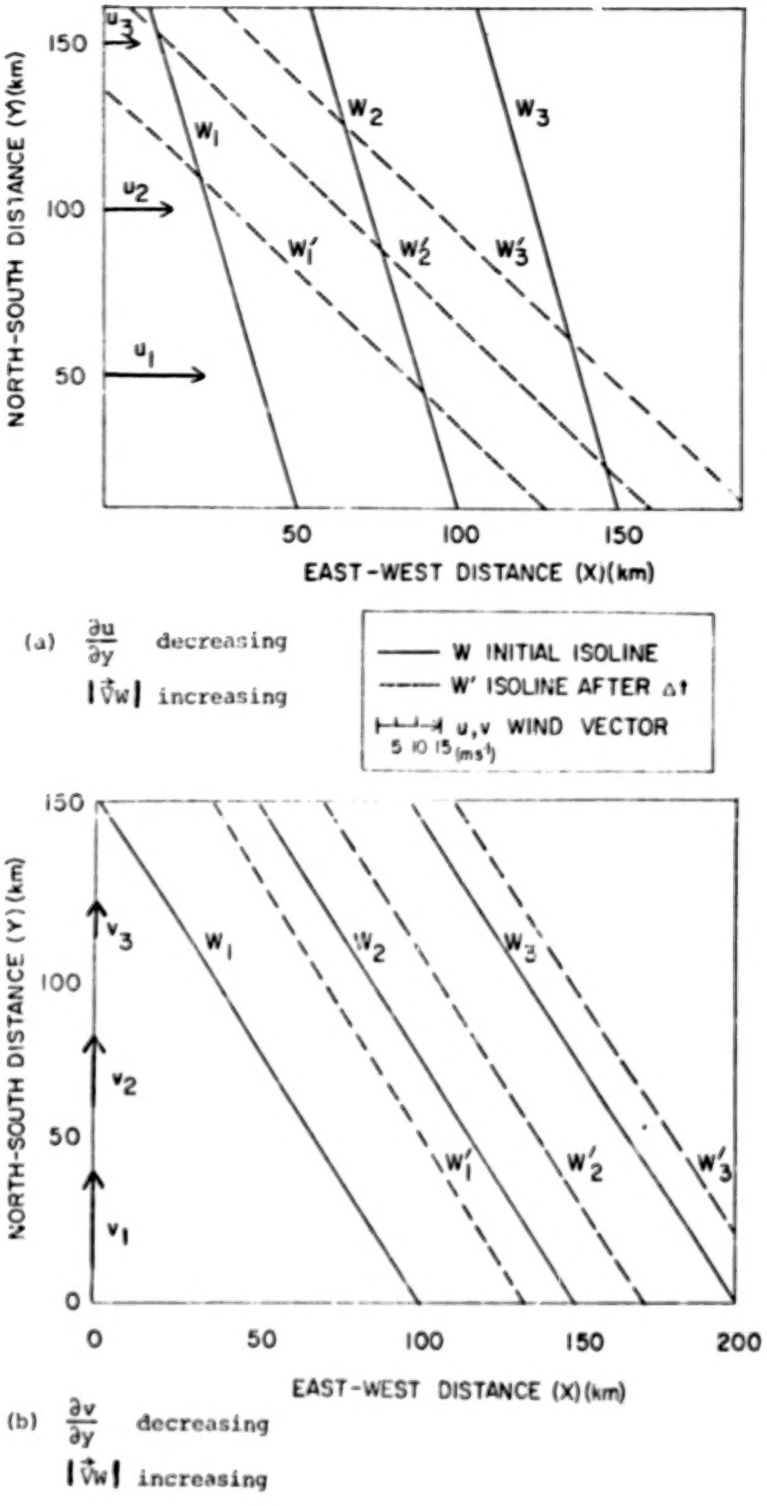


Fig. 7. Schematic illustration of isoline concentration by (a) shear and (b) confluence.

with time. Similarly, the horizontal temperature gradient is related to the vertical wind shear and differential temperature advection will be reflected in adjustment of the thickness field. Both wind shear and thickness differences have been included. The Laplacian of thickness advection should be a way of locating centers of strong differential temperature changes which, in a subsequent time interval, could be related to thunderstorm development. Convective instability is important to thunderstorm development (Koch, 1975). This type exists in the atmosphere in those layers where θ_e decreases with height. There are three parameters in which a finite difference version of this term are included. The last of these is static stability discussed by Paine and Kaplan (1974). Finally, standard parcel stability measures and surface values of temperature and potential temperature were used.

3) Trigger mechanism

Many days occur when sufficient moisture and instability are both present and yet there are no thunderstorms. A trigger mechanism is needed to release the instability and latent energy. Usually, this trigger is manifested in vertical motion, so that we need to find a lifting mechanism. Terrain-induced vertical motion is included as a predictor combined with surface velocity divergence. The vorticity field at the surface measured by the vertical component of the curl of the surface wind field or indirectly through the pressure Laplacian is another potential uplift mechanism through convergence which it induces. Fronts are frequently associated with thunderstorms. A front can be identified through the wind, temperature, moisture, and pressure fields. Temperature, moisture, pressure gradients and the advections

of θ_e and P were included as parameters. Measures of vertical motion can be obtained in only a crude way from data at just five levels in the atmosphere. Both integrated divergence (sums of divergence for three levels) and integrated moisture divergence were included as predictor parameters. Differential vorticity advection (DVA) is included as a parameter since it together with the Laplacian of thickness advection, are the two terms in the ω-equation (Holton, 1972). The meridional wind component at 500 mb is a measure of the strength and/or proximity of an approaching trough if a general west to east wave motion exists. Vorticity advection and vertical motion usually ensue. The v-component sum at 850 and 500 mb measures the degree to which the wind is in-phase at these two levels east of a trough. The more out-of-phase, the lower this sum would be; therefore, one would be looking at a measure of the baroclinity of the lower atmosphere. A negative correlation of this parameter measured at 1200 GMT with thunderstorms later in the day would be expected. Finally, an increased tendency for cyclogenesis at the surface may be associated with general uplift and, therefore, a trigger mechanism for subsequent thunderstorms. The time change of the Laplacian of the surface pressure, $\frac{\partial}{\partial t}(\nabla^2 p)$, is one such indicator.

The last trigger shown in Table 8 is a binary radar parameter. Any MDR code (two or greater) during the time period 1700 to 1800 GMT for each grid point was coded as one; otherwise, zero was assigned. In this way a one represents any precipitation occurring near the time the forecast is to be made. Usually, when other conditions are right, any precipitation at this time of the morning either maintains its intensity

by propagating within the predictand area when the code is already greater than four, or develops into a thunderstorm in the subsequent 2 to 5 h. This predictor is the only direct measure of vertical motion or trigger among all predictor parameters. Of course, some of the parameters could contribute to more than one condition for thunderstorms. Consider, for example, the discontinuity function, |D|; while listed under moisture, it might also be discussed in conjunction with the dry line and frontogenesis or a trigger term. Similarly, the Laplacian of thickness advection is a term in both the ω-equation and Petterssen's development of surface vorticity tendency. It could be shown with the trigger terms as well.

c. Subdivision of original data

The total data set consists of parameters calculated at each of 256 grid points for 240 days. However, for reasons discussed in Section 3, not every grid-point was used. There were a total of 7680 observations possible in the data set used for subsequent statistical analysis. However, an observation which contained any missing element was not used. The data were then subdivided into groups as shown in Fig. 8.

1) Developmental and test

Subdivision of the original data set into developmental and test groups was necessary so that some type of quality measure or verification could be obtained. Every third day is considered to be independent for temperature (Panofsky and Brier, 1958). Therefore, data in every third day (day one, day four, day seven, ...) were used as a test sample. The developmental sample included data in all other days. As far as thunderstorms were concerned, the assumption of independence was

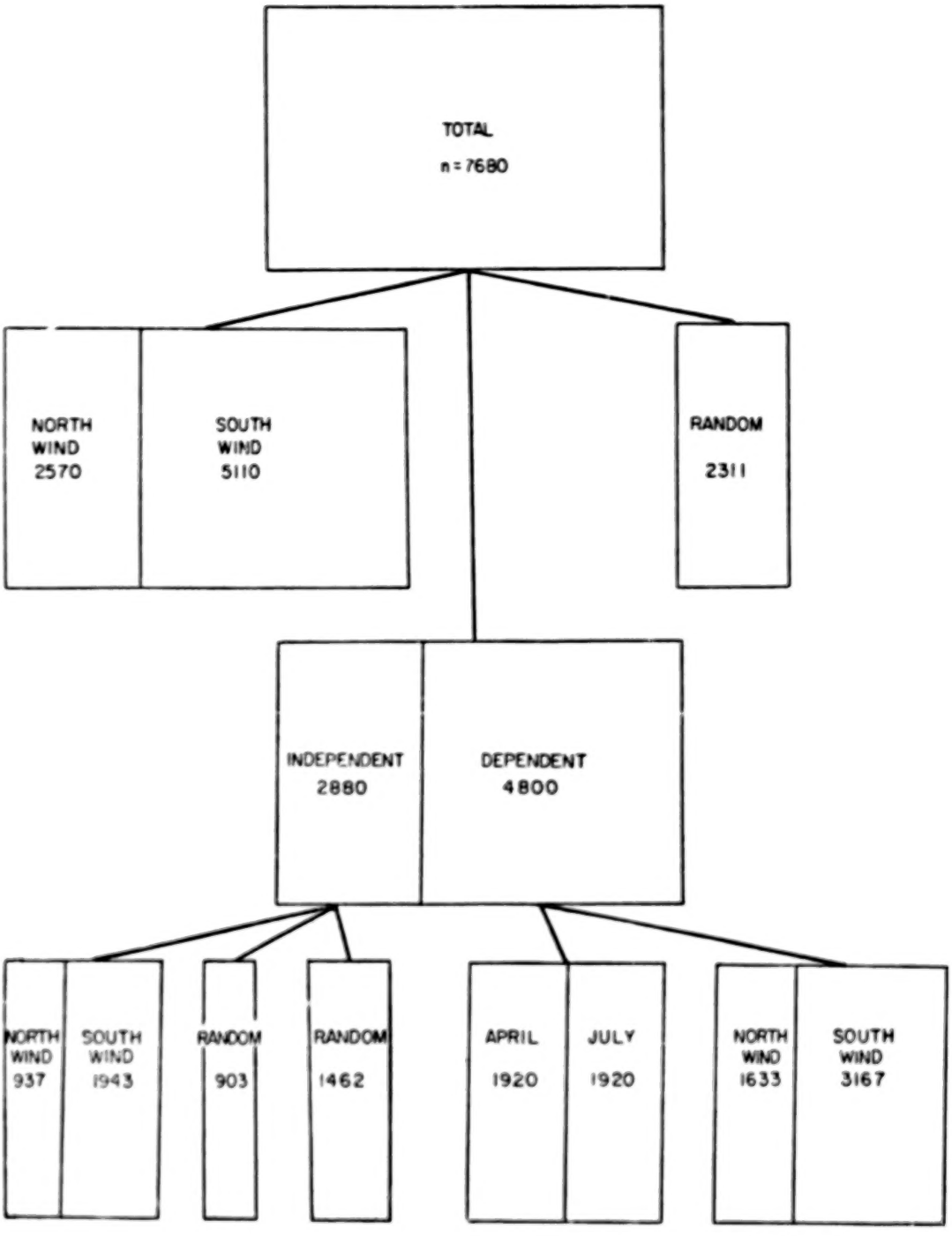


Fig. 8. Subdivision of total data set.

examined for a few grid points in the test sample. Ones were not observed for two consecutive periods (day one and day four, for example). The developmental sample was, therefore, considered to be the dependent sample; the test data set was the independent sample. Equations were developed from statistical models applied to the developmental sample and tested on the test sample.

2) North wind and south wind

Thunderstorms are observed to develop and behave somewhat differently in different types of synoptic situations or in different air masses (Purdom, 1975). Subdivision by air mass may give better results in this type of statistical analysis where the sample includes several thunderstorm seasons for a large area. Partitioning by air mass was not directly possible with the historical data available. However, a division of data by surface wind component at 1800 GMT was considered to be a fair substitute. Consequently, the data were divided into north wind and south wind sets depending on whether or not the surface wind had a northerly or southerly component at 1800 GMT, respectively. Separate regression analyses were performed on each subset.

3) April and July

All days and observation points in April for the two years of data were combined. The same was done for July. Again, analyses were performed within each data set to determine differences, if any, in spring and summer predictors.

4) Random sample

Samples were chosen by random-number generators so that developmental samples contained nearly the same number of occurrences and

nonoccurrences. The unequal natural frequencies of thunderstorms versus no thunderstorms create problems in regression analysis when the dependent variable is binary. These problems were discussed in Section 2. Results of application of the various statistical techniques outlined in Section 2 to subsets discussed here are presented next.

5. RESULTS

First, comparisons of results for different predictand times will be presented. The types of predictors selected and the order in which they were included in the model will be discussed next for all data subdivisions. Following this will be the importance of surface versus upper-air parameters to the prediction of thunderstorms. A discussion of the maximum R^2 or variance reduction achieved will follow. Next, performance of the equations applied to an independent data set will be presented followed by comparisons with results of other investigators. Results of a principal component analysis will be presented next. Last will be a discussion of the utility of these equations in an operational environment.

a. Forecast time intervals

Regression models were tested with fixed numbers of independent variables and three time combinations of the dependent variable: 1900-2000 GMT, 2000-2300 GMT, and 2300-0200 GMT. Random samples were chosen so that p was nearly the same. The R^2 decreased, as expected, when the time interval between observations and forecasts lengthened; however, the occurrence frequency of the predictand was only 9.3% in the first period. With so few occurrences, this equation would likely deteriorate[9] when applied to independent data. In other words, there

---

[9]"Deteriorate" means that probabilities of thunderstorms produced by the linear equation developed from data in a dependent or developmental sample would not correspond well with observed frequencies of occurrence when these equations are used on an independent sample of data.

would not have been enough different thunderstorm-producing environments included in the sample. Also, extrapolation of existing radar echo patterns would seem to be a more promising technique for those 1- to 2-h forecasts. Similarly, it is not likely that observed features of the atmosphere early in the morning would adequately reflect ingredients for the occurrence of thunderstorms late in the afternoon. Consequently, only the 2000-2300 GMT period was included in all further analyses.

b. Predictor selections

The order of selection and specific predictors selected by a stepwise, variable selection technique are shown in Table 9 for different groups of data. Only the first six of many predictors offered as candidates are shown. No matter how the data are divided, the three variables consistently selected include a combination of moisture and trigger terms; the next several invariably include a measure of atmospheric instability through either stability indices or linear combinations of vertical temperature and moisture parameters. The first four-to-five variables include all the synoptic-scale conditions for intense convection. Therefore, it is not surprising that more than 85% of the total variance explained by the regression model is accounted for by the first five variables.

In the case of the north wind and all other subsets except south wind, the single most important predictor was the surface mixing ratio. The presence of precipitation (MDRP) near 1800 GMT was most important for the south wind data. In the area chosen for this study, a south wind implies the presence of maritime tropical air which contains

Table 9. Variables selected, cumulative R^2, and sign of coefficients for a stepwise selection procedure and different data subsets.

	Order																	
Data	1			2			3			4			5			6		
Total	W	0.108	+	MDRP	0.181	+	MDIV	0.200	-	$T-T_{d_7}$	0.209	-	CSIL	0.216	-	θ_{e_7}	0.220	-
Random 18% of no-thunderstorm days	W	0.185	+	TTI	0.234	+	MDRP	0.266	+	MDIV	0.280	-	$T-T_{d_7}$	0.288	-	θ_{e_7}	0.296	-
North wind	W	0.166	+	MDRP	0.220	+	MDIV	0.250	-	$T-T_{d_7}$	0.254	-	θ_{e_7}	0.259	-	$T-T_d$	0.266	+
South wind	MDRP	0.105	+	W	0.160	+	MDIV	0.176	-	$T-T_{d_8}$	0.191	-	$T-T_{d_7}$	0.197	-	θ_{e_7}	0.202	-
April dependent	W	0.121	+	MDRP	0.187	+	MDIV	0.216	-	V_s	0.224	+	LDI	0.228	+	$T-T_{d_7}$	0.232	-
July dependent	W	0.130	+	MDRP	0.201	+	MDIV	0.219	-	TTI	0.229	+	VSUM	0.239	-	$T-T_{d_7}$	0.244	-

considerable moisture. Therefore, a trigger mechanism identified by MDRP would be an important parameter contributing to thunderstorms, given that moisture is already present.

For the April and July subsets the first three predictors are the same. During April, a 500-mb trough (v-wind component at 500 mb) and concentration of moisture gradient at the surface through the deformation field of the wind are the next most important parameters. This latter predictor can be interpreted to represent the location of the surface dry line which is recognized as a favored region for severe weather (Miller, 1972). In the spring, surface winds are stronger and gradients more intense than in summer. Therefore, one would expect these quantities to be reflected more in the synoptic data which are utilized. During July stability measured by the Total-Totals Index is the fourth predictor chosen. This development is reasonable owing to the weaker winds in the summer.

In a separate analysis, four different time changes were computed for five surface variables. These were the 1-h, 3-h, 6-h and 3-h change in the 3-h time change for the following: θ_e, MDIV, WTS, θ_eA, and ∇^2P. When these were used as candidate predictors in the stepwise selection procedure, they were not chosen among the top five predictors. Also, when time derivatives were selected, the 3-h and 6-h changes were chosen before 1-h changes. One possibility for this result is that the original spacing of surface data and analysis procedures restricts the amplitudes of resolvable features. Six-hour features are more likely to have the larger amplitudes which can trigger intense convection later in the afternoon. More work needs to be done in this area.

The signs of regression coefficients are as expected when other variables are included in the model. For example, the sign of the temperature coefficient is interpreted as the change in predictand for a unit change in temperature while holding constant all other variables in the model at that time. The negative sign indicates that given that surface moisture (among other things) already is present, then thunderstorms occur with lower temperatures or when the air is more nearly saturated. The total correlation coefficient for temperature shown in Table 10 indicates a positive correlation of temperature and thunderstorms when all other variables are ignored.

c. Importance of surface versus upper-air parameters

Regression models were utilized with stepwise procedures for surface variables and upper-air variables separately. Results are summarized in Table 11. Surface parameters alone in linear combination accounted for 15% of the total variance ($R^2 = 0.150$), whereas upper-air parameters accounted for only 13.4%. When both sets were used together, however, the best results were obtained; R^2 improved to 0.197. Both timeliness and spatial resolution contributed to this result. Surface data were available at 1800 GMT, 2-5 h before thunderstorm occurrence as opposed to upper-air observations at 1200 GMT. Also, surface stations are spaced about 120 km apart compared to 370 km for upper-air reports. Space derivatives, which are used extensively as parameters, are, therefore, more nearly represented by finite differences in the case of the former. Even though the upper-air predictors were old and contained poor spatial resolution, when combined with surface predictors, they produced a 30% improvement in R^2. It appears

Table 10. Linear correlation coefficients of selected predictors with the occurrence of thunderstorms during the period 2000-2300 GMT.

Predictor	Time of Observation (GMT)	Correlation Coefficient	Significance Probability level
θ_e	1800	0.280	0.0001
ω_{TS}	1800	-0.154	0.0001
MDIV	1800	0.173	0.0001
$\theta_e A$	1800	0.103	0.0001
LP	1800	0.079	0.0001
ζ	1800	0.099	0.0001
$\|D\|$	1800	0.050	0.0001
$\|\vec{\nabla}w\|$	1800	0.041	0.0006
$\vec{\nabla}^2 w$	1800	-0.082	0.0001
CSIM	1200	-0.275	0.0001
CSIL	1200	-0.230	0.0001
KI	1200	0.289	0.0001
TTI	1200	0.251	0.0001
STSI	1200	-0.274	0.0001
UWSH	1200	-0.130	0.0001
DVA	1200	0.008	0.5173
LTHA	1200	0.052	0.0001
DTA	1200	0.051	0.0001
IDIV	1200	-0.040	0.0008
IMDIV	1200	-0.041	0.0006
$\|\vec{v}_5\|$	1200	-0.128	0.0001

Table 10. (Concluded)

Predictor	Time of Observation (GMT)	Correlation Coefficient	Significance Probability level
DTH	1200	-0.016	0.0001
θ_{e_8}	1200	0.283	0.0001
θ_{e_7}	1200	0.220	0.0001
W_8	1200	0.311	0.0001
W_7	1200	0.233	0.0001
$(T-T_d)_7$	1200	-0.174	0.0001
$(T-T_d)_8$	1200	-0.213	0.0001
v_5	1200	0.090	0.0001
W	1800	0.328	0.0001
VSUM	1200	0.113	0.0001
u	1800	-0.043	0.0001
v	1800	0.050	0.0001
T	1800	0.166	0.0001
$T-T_d$	1900	-0.190	0.0001
MDRP	1735	0.324	0.0001

Table 11. Summary of statistics for regression analyses with surface and upper-air predictors.

Total Sample Size	Occurrence Frequency	Data	Max R^2	Number of Predictors	Mean Squared Error
7492	17.9	Surface	0.150	11	0.125
7492	17.9	Upper air	0.134	16	0.128
7492	17.9	Surface and upper air	0.197	24	0.118
7125	17.9	Upper air, surface and MDRP	0.243	20	0.114

that poor as they are, these predictors fill an important gap in identifying those observed features of the atmosphere which are subsequently related to intense convection. Surface data alone give little indication of the potential stability of the atmosphere. It is this ingredient which is added by including upper-air parameters. The dew-point depression at the 700-mb level is the first upper-air predictor included by the stepwise procedure. Also, it is the third parameter following low-level moisture and moisture divergence. Stability alone, however, gives inadequate information for predicting subsequent thunderstorms. From Table 10 (p. 57) it is seen that the highest correlation coefficient between the predictand and any single stability measure is 0.289 for the K index. Several other variables such as W, the radar predictor (MDRP), and equivalent potential temperature differences exhibit higher correlations.

d. Quality of fit of the regression model

While the conditions for thunderstorms are known with some confidence as far as synoptic data are concerned, there is little confidence in determining these conditions from the study data. For example, stability can be obtained from the vertical structure of temperature and moisture profiles. When a limited sample of these data at a few fixed levels in the troposphere comprise our measures, only approximations to the stability can be made. There is a number of these approximations depending on levels, variable combinations, and physical assumptions (parcel method, layer method, etc.). Similarly, the trigger mechanism must be inferred since vertical motion, the usual trigger, is not one of the observed variables. Finally, the parameters contributing to many thunderstorm occurrences exist on a scale much smaller than we can resolve with our data. Thunderstorms have been observed to occur at boundaries and intersections of pressure discontinuities (gust fronts) caused by previous cells (Purdom, 1974). Similarly, they have been observed to develop in the afternoon in areas which were void of clouds that morning (Weiss and Purdom, 1974). The influence of the sea breeze is illustrated by the frequency distribution of thunderstorms along the Gulf Coast and Florida (Scoggins, 1976). Small-scale convergence induced by gravity waves (Wave CISK) appears to be important to intense convection from theoretical considerations as well (Raymond, 1976). Even diffusion in a two-constituent medium might be a trigger (Schaefer, 1975). A consequence of the foregoing discussion is reflected in the overall low R^2 or relatively small amount of variance of thunderstorm occurrences that can be explained by the linear combination of synoptic

parameters.

Table 12 contains the maximum R^2 for a specific number of predictors in each data subset. The regression for the random sample of no-thunderstorm observations produced the highest R^2, 0.332, most likely because

Table 12. Summary of statistics for regression analyses with different data subsets.

Total Sample Size	Occurrence Frequency (%)	Data	Max R^2	Number of Predictors	Mean Squared Error
7125	17.9	Total	0.243	20	0.114
2203	40.7	Random 18% of no TSTM days	0.332	18	0.163
2376	13.9	North wind	0.284	13	0.086
4750	20.6	South wind	0.238	21	0.125
1837	8.1	April dependent	0.255	14	0.056
1759	25.4	July dependent	0.279	16	0.138

the total number of observations decreased. The equation for predicting thunderstorms which developed between 2000 and 2300 GMT following surface wind with a northerly component at 1800 GMT accounted for 28.4% of the total variance, whereas the south wind equation accounted for 23.8% though some of this difference would be due to the larger occurrence frequency in the south wind data. Further, the north wind equation did its job with a fewer number of predictors.

The R^2 for the April and July data are based on fewer observations,

but it is interesting to note that the R^2 for April (0.255) is lower than that for July (0.279) even though the frequency of thunderstorm occurrence is much higher in July.

The mean squared error (MSE) of the regression analyses continued to be reduced as more variables were added to the model. This indicates that the exact synoptic-scale measures of the conditions for thunderstorms were not available, or the parameters did not truly represent these conditions. This result is not surprising if one considers the crudeness of our measures of atmospheric structure in terms of limited horizontal and vertical resolution, the untimeliness of the upper-air measurements (8-11 h before thunderstorm occurrence), and limitations imposed by the specific observed variables from which parameters were computed.

Another way to evaluate quality is to consider how well predicted probabilities represent actual frequencies of occurrence of thunderstorms. Predicted probabilities in 10% increments were generated for several different data subdivisions. These are shown in Fig. 9. In general there was a slight tendency to overpredict the observed probability at low probabilities and underpredict for probabilities above 0.6. This seems to be consistent with our natural bias in subjectively-derived probabilities. Underprediction at high frequencies of occurrence can be explained by the decreased slope of the regression plane owing to the many more non-occurrence observations compared to thunderstorm occurrences (see paragraph e). No explanation, however, is apparent for the overprediction.

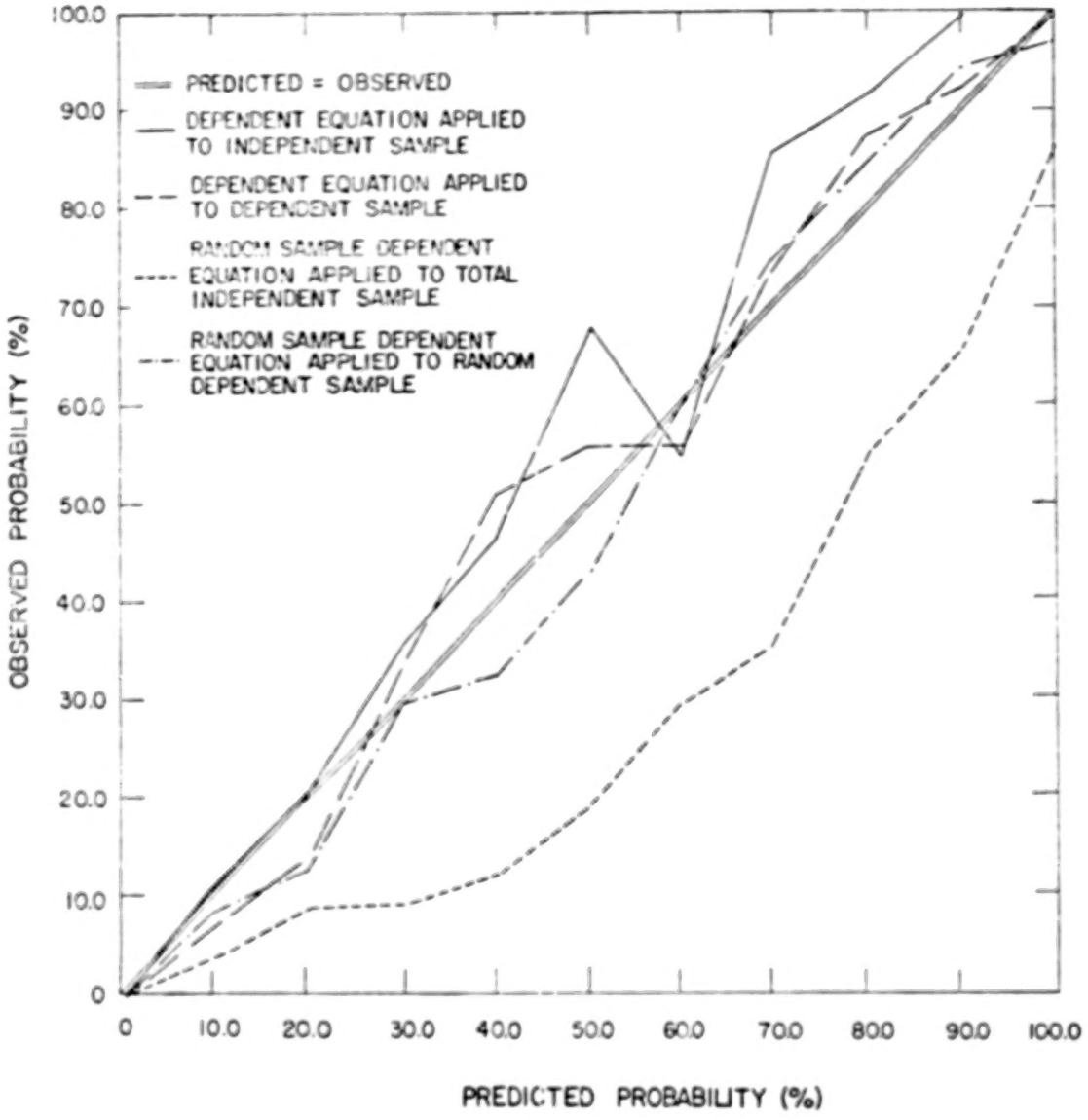


Fig. 9. Relation between predicted and observed probabilities of thunderstorm occurrences for various data subdivisions.

e. Performance on a test data sample

A further measure of the quality or goodness of the regression models is how well the equations perform on an independent data sample. Equations developed from a dependent sample were applied to independent data. Furthermore, a threshold of predicted probability was chosen and a contingency table of counts for predicted and observed yes and no cases developed as follows:

Observed \ Forecast	Yes	No	Sum
Yes	A	B	S_1
No	C	D	S_2
Sum	S_3	S_4	T

From such a table some typical discriminates can be examined such as the overall percent correct, (A+D)/T)100; the percent of correctly forecast observations of an occurrence, called prefigurance, $(A/S_1)100$; the percent of correctly observed forecasts of occurrence (postagreement), $(A/S_3)100$; Threat Score, A/A+B+C (Charba, 1977) called critical success index by Donaldson et al., (1975); skill score $[(A+D) - S_3S_1 + S_4S_2)/T]/[T - S_3S_1 + S_4S_2)/T]$, discussed by Brier and Allen (1952); and V-score, $V = (AD - BC)/(A+B)(C+D)$, presented by Dobryshman (1972) and discussed by Woodcock (1976). The threat score, skill score, or percent correct cannot be interpreted to measure relative merits of each subdivision of the original data because each is a function of the observed probability of occurrence (called trial conditions by Woodcock (1976)). These probabilities change as the threshold of predicted probabilities for classification purposes changes. Table 13 contains example contingency

Table 13. Contingency tables for 25 no-thunderstorm forecasts shifted to the yes forecast column in different observed proportions. Threat score and skill score are also shown.

(a) Original proportion (20% forecast yes)

Observed \ Forecast	Yes	No
Yes	67	55
No	108	656

TS = 0.291

SS = 0.340

(b) 10% proportion (3 yes; 22 no)

Observed \ Forecast	Yes	No
Yes	70	52
No	130	634

TS = 0.280

SS = 0.318

(c) 15% proportion (4 yes; 21 no)

Observed \ Forecast	Yes	No
Yes	71	51
No	129	635

TS = 0.282

SS = 0.326

(d) 20% proportion (5 yes; 20 no)

Observed \ Forecast	Yes	No
Yes	72	50
No	128	636

TS = 0.288

SS = 0.330

(e) 30% proportion (8 yes; 17 no)

Observed \ Forecast	Yes	No
Yes	75	47
No	125	639

TS 0.304

SS = 0.356

(f) 49% proportion (12 yes; 13 no)

Observed \ Forecast	Yes	No
Yes	79	43
No	121	643

TS = 0.320

SS = 0.386

tables where a fixed number of no-thunderstorm forecasts (in this case 25) are shifted to the yes column for different proportions of observed yes and no cases. This is exactly what is done when the threshold probability is lowered. One can see that the threat score (TS) or skill score (SS) exceeds the original values only after the proportion within the observed categories exceeds the original forecast probability. They appear to be unsuitable for a goodness measure. The overall percent correct also is not very meaningful because of the many days when no thunderstorms occur. The V-score is least affected by trial conditions but also involves the No-No entry. Therefore, our discussions will focus primarily on the prefigurance and postagreement percentages. Table 14 contains the above discriminates for each data subdivision.

One can obtain an indication of the deterioration of the equations by looking at the decrease in any of the discriminates but, in particular, the V-score between the developmental and test samples. For example, the mean V-score for the total developmental sample is 0.454. The mean for the total test sample is 0.390. The lower score means poorer performance.

Thunderstorms appear to be more predictable from synoptic parameters when the surface wind has a northerly component at 1800 GMT. Such an implication is indicated by the greater V-score, prefigurance, and postagreement percentages for the north wind equation when tested on the independent sample compared to similar statistics for either the total equation or that for the south wind subdivision. There are several explanations. First, thunderstorms frequently develop behind a shallow surface cold front (north wind component) in the area of

Table 14. Summary of contingency tables.

Observed Frequency (%)	Forecast Frequency (%)	Threshold	Prefigurance (%)	Postagreement (%)	Percent Correct	Skill Score	Threat Score	V score	Sample to which applied
				Total Sample					
19.4	31.4	.22	65	40	74	.337	.329	.417	Test
19.4	25.7	.25	59	44	78	.362	.335	.407	Test
19.4	20.2	.28	51	49	80	.379	.335	.385	Test
19.4	17.1	.30	45	51	81	.350	.316	.350	Test
17.0	26.8	.25	69	44	85	.418	.369	.513	Developmental
17.0	18.1	.30	54	51	83	.427	.358	.438	Developmental
17.0	15.5	.32	50	54	84	.426	.351	.411	Developmental
				Random Sample					
60.4	62.4	.46	78	80	74	.455	.652	.451	Random test
59.5	56.7	.50	57	68	75	.474	.638	.500	Random test
57.1	63.2	.50	85	77	76	.509	.672	.500	Random developmental
19.4	31.9	.42	64	31	61	.243	.295	.399	Test
19.4	40.2	.50	76	37	70	.316	.329	.445	Test
19.4	20.8	.65	52	48	80	.381	.337	.392	Test
				North Wind					
13.8	23.2	.25	66	39	81	.381	.323	.491	Test
13.8	20.5	.28	64	43	83	.417	.345	.503	Test
13.8	18.5	.30	59	44	84	.410	.336	.470	Test
13.8	14.1	.34	47	46	85	.374	.300	.378	Test
13.8	11.3	.37	39	47	85	.339	.267	.314	Test
13.5	22.3	.25	74	45	84	.466	.385	.596	Developmental
13.5	17.5	.30	65	59	87	.494	.399	.556	Developmental
13.5	13.8	.34	59	57	89	.513	.408	.518	Developmental
13.5	12.2	.37	52	58	88	.483	.379	.464	Developmental
				South Wind					
22.0	26.3	.25	57	48	77	.372	.354	.399	Test
22.0	23.0	.27	54	52	79	.389	.357	.396	Test
22.0	17.0	.30	45	58	81	.385	.336	.354	Test
19.0	29.8	.25	69	44	77	.398	.366	.486	Developmental
19.0	25.6	.27	63	47	79	.409	.367	.464	Developmental
19.0	20.6	.30	56	51	82	.423	.367	.438	Developmental
19.0	16.1	.33	47	55	83	.404	.340	.382	Developmental

this study. In these situations the storms are usually connected with a synoptic-scale vertical motion field that results from positive vorticity advection due to a short-wave trough aloft, given that moisture and potential instability exist. Storms also can develop along a surface cold front which trails an active squall line. In these cases as well, the surface winds behind a southeastward moving squall line are likely to have a northerly component 2-5 h before the occurrence of the cold-front cells. Finally, we can distinguish between thunderstorms in continental air masses where surface winds are from the North and maritime air masses with southerly winds. Thunderstorms occurring in the maritime air are more frequently classified as convective, air-mass thunderstorms (Beers, 1945). The trigger mechanism for releasing the instability usually present is less detectable from synoptic data. Mesoscale or even smaller discontinuities may exist and contribute to the trigger. These elude detection from the data in this study.

When applied to a random dependent sample (in other words how well can the linear function discriminate between thunderstorms and no thunderstorms within the dependent sample which only includes 17% of all no observations), the equation produced prefigurance and post-agreement percentages of 85 and 77, respectively, although the overall percent correct was down to 76 (Table 14, p. 67). The deterioration when applied to a random independent sample was not large. For a threshold of 0.46, 78 and 80% were obtained for the prefigurance and postagreement, respectively. When the equation developed from the random dependent sample (17% of no-thunderstorm observations) was applied to the total

independent sample (as opposed to the random independent sample), the prefigurance-postagreement percentages were not as high.

It is not clear why the equation from the random dependent sample deteriorates so little when applied to the independent sample. One possible explanation could be due to the binary nature of the dependent variable and unequal distributions of occurrences and nonoccurrences. Figure 10 shows how the influence of the nonoccurence observations.

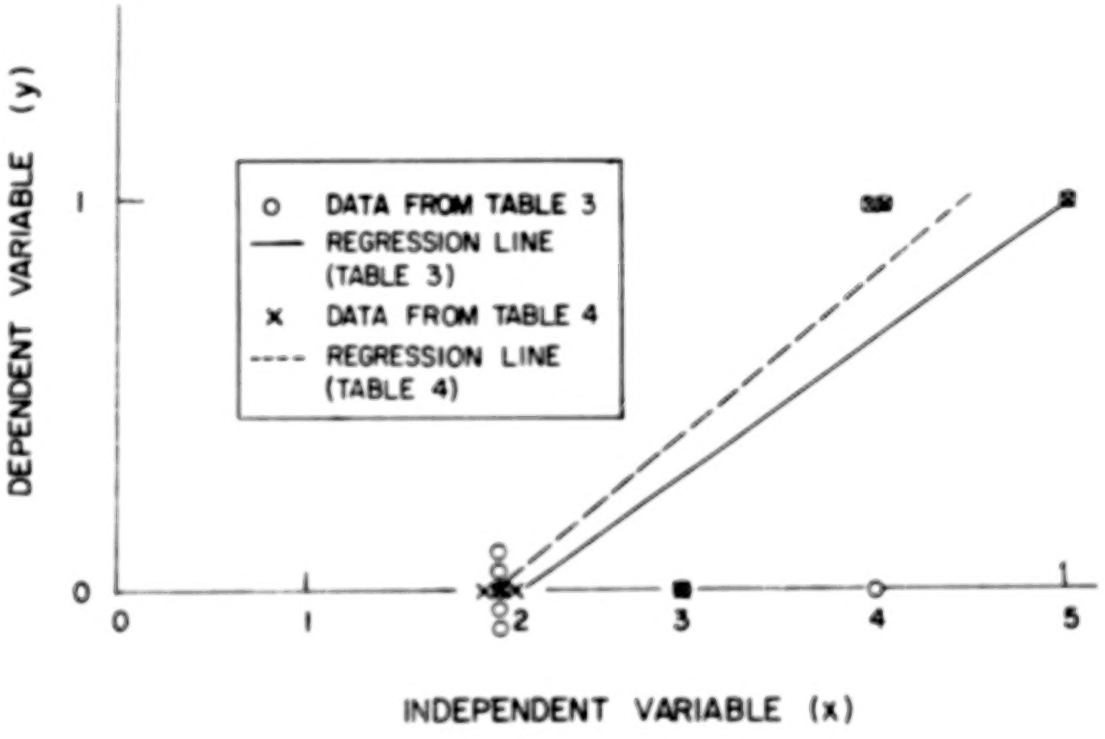


Fig. 10. Regression lines for data in Tables 3 and 4.

actually decreases the slope of the least squares estimate of a regression line. By eliminating the no's so that the proportions are more nearly equal, the slope is increased. This increase can be visualized as an increase in the discriminating ability of the independent

variables. As the slope of the regression line increases, a small change in x would produce a large change in the predicted probability (Y) if the linear function were to be used in a predictive fashion.

Several attempts were made to accomplish the same result by using critical values of predictors. These values were selected from frequency distributions of the predictand and leading predictors. One such frequency distribution is shown in Fig. 11. There a cut-off would be 5 g kg^{-1} for the surface mixing ratio. Others were chosen similarly and used in conjunction (logical and) and disjunction (logical or) operations. An example of the latter would be as follows: If $W < 5$ g kg^{-1} or $\theta_e < 317$ K or KI < -8 or $W_8 < 5$ g kg^{-1}, then delete this observation (hopefully it will be a no-thunderstorm observation). In fact, the above statement provided the best results which could be obtained. Frequency of occurrence was increased only 7% and R^2 changed from 0.260 to 0.247 for a stepwise procedure.

f. Comparison with other results

Except for the work of Charba (1977) there are no other results which are directly comparable. Charba has published his results for a similar statistical technique (step up) and for 2- to 6-h forecasts of thunderstorms (defined similarly from MDR data). His research area includes most of the eastern United States, and predictand area is about 80 km on a side. However, Charba used combinations of radar observations at 1735 GMT, radar climatology, surface observations at 1500 GMT, and upper-air forecasts valid at 2100 GMT from a limited-area, fine-mesh model (Howcroft and Desmarais, 1971) as predictors.

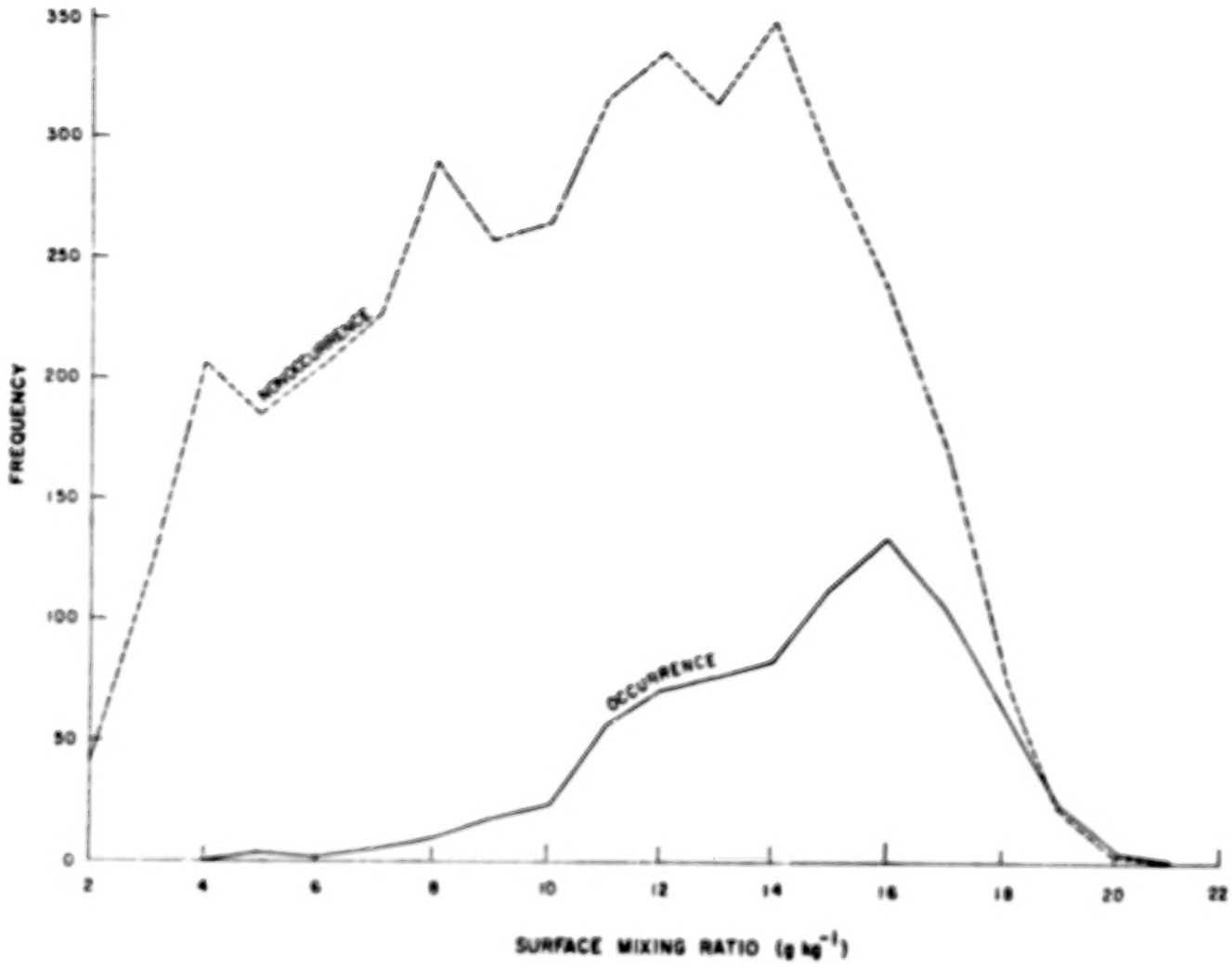


Fig. 11. Frequency distributions of occurrence and nonoccurrence of MDR≥4 for different values of the surface mixing ratio.

If we exclude radar predictors, his top four were (1) a modified[10] K index, (2) moisture divergence at the surface, (3) modified Total-Totals Index, and (4) 500-mb wind speed. These compare favorably with the moisture, divergence, and stability parameters from observations in this study. The observed frequency of thunderstorms in Charba's work was 10% compared to 17% here. One should see an increased R^2 from this influence in Charba's result counteracted to some extent by a reduction in R^2 due to the smaller forecast area. The net result was an R^2 of 0.282 in Charba's scheme compared to 0.284 in the case of our north-wind equation. In addition, Charba's predicted probabilities were similar overall to those in this research.

Some knowledge is required of how forecasters subjectively predict thunderstorms in an operational environment. Unfortunately, there are no statistics which would exactly correspond to the areas, times, and procedures used here. In fact, any verifications of thunderstorm forecasts with different lead times are difficult to find. One set of data was available for 14 base weather stations in or near the area of Fig. 4 (p. 32) during the June, July, and August 1976 period. These data consist of warnings issued by forecasters of impending thunderstorms. The number issued and the number verified with a lead time is summarized in Table 15. Thunderstorms which occur less than 1 h from the forecast

---

[10] Modified in this context means that surface observations of temperature and dew point at 1800 GMT were averaged with a forecast temperature and dew point at 850 mb.

Table 15. Contingency table of observed and forecast thunderstorms for 14 base weather stations near the area outlined in Fig. 4.

		Forecast: Yes	Forecast: No	
Observed	Yes	78	90	46.4%
	No	167		
		31.8%		

time are counted as misses. For example, a warning for thunderstorms issued at 1700 GMT valid for the period 1900 to 2300 GMT would be a hit if a thunderstorm were observed at the station or within the base environment after 1900 GMT. Otherwise, it would be a miss. The base environment is usually about a 10-km radius of the station but may vary up to 45 km. This is still considerably smaller than the forecast area of a square 138 km on a side used in this study and, therefore, should reflect poorer performance. On the other hand, a 1-h lead time is allowed for verifying the weather warnings, whereas the lead time is 2 h for the statistics in Appendix B. This may compensate to some extent for the smaller area. Many other differences exist between these statistics and those presented in Appendix B so that comparisons are difficult. The very definition of thunderstorms is different. An MDR code of four or greater was used in this research. The weather station

used their observation log or the radar in a qualitative sense. Also, the issue time from the weather station was not constrained to 1800 GMT. Finally, the period for the base weather station included a year, 1976, and month, August, that were not available in this study. Nevertheless, the low prefigurance and postagreement percentages of 46% and 32%, respectively, seem to be typical of a forecaster's performance at this difficult tack.

Though there is little confidence in comparisons of verification measures applied to data of this nature, there are a combination of encouraging signs which lead to a conclusion that observations of key parameters in linear combination can provide useful forecasts of thunderstorms in areas of about 8350 km^2 for periods of 2- to 5-h. First, parameters selected by statistical methods provide the ingredients for subsequent thunderstorms which have been deduced from many years of experience. Secondly, the equations do not deteriorate when applied to independent samples. Further, contingency tables produced from equations for many different threshold probabilities provided higher prefigurance and postagreement percentages than those from a table of actual performance. Also, predicted probabilities from the equations represent actual occurrence frequencies. Finally, these results are very similar to those from an operational program where forecast model predictors had been used. Results of a principal component analysis are discussed next.

g. Dimensionality

As stated earlier, eigenvectors of the independent-variable matrix

that consist of sums of squares and cross products, (X'X), can be interpreted to represent the part of the total variance accounted for by the given linear combination of variables where the eigenvector elements are the weights or coefficients. If it turns out that the first few components account for some large percentage of the total variance as shown by the cumulative portion of the eigenvalues, then it can be assumed that there is evidence of the true dimensionality of the original set of variables or that there is an indication of the total number of hypothetical, latent variables needed to describe the structure of the original variables. This is another way of quantifying the degree of intercorrelation among the x's. These eigenvectors for different subsets of the (X'X) matrix are shown in Table 16. Also shown are the associated eigenvalues and cumulative portion of the total variance which is accounted for by each successive eigenvector.

In the case of moisture parameters, we can account for nearly 90% of the total variance in all moisture parameters by using the first five components (the five largest eigenvalues). We can account for 50% of the total with just two. The variables which seem to be most important, according to the sum of the first two eigenvector coefficients, are surface, 850-, and 700-mb mixing ratio, equivalent potential temperature at the surface, and dew-point depression at 700 mb. It is not surprising that among these are the leading parameters selected by the stepwise regression procedure.

Stability parameters have fewer dimensions as shown by the eigenvectors. Just one principal component accounts for 59% of the total variance. The 90% point is reached with only four eigenvectors. Among

Table 16. Eigenvectors and eigenvalues for moisture, stability and trigger parameters.

(a)	Moisture Parameters					
	Eigenvectors					
Parameter	1	2	3	4	5	6
θ_e	0.442	0.372	0.044	0.156	0.035	0.304
$\lvert\nabla w\rvert$	0.080	0.259	0.585	0.400	-0.004	-0.649
$\nabla^2 w$	-0.095	-0.213	0.543	-0.094	0.714	0.341
$\theta_e A$	0.027	0.192	-0.577	0.187	0.693	-0.334
$\bar{w}_8$	0.489	0.061	0.118	-0.326	0.077	-0.056
w_7	0.403	-0.408	-0.016	0.371	-0.006	0.095
$(T-T_d)_7$	-0.255	0.630	0.101	-0.312	0.026	0.113
$(T-T_d)_8$	-0.331	0.196	-0.003	0.653	-0.045	0.415
w	0.460	0.327	-0.055	0.068	-0.022	0.249
Eigenvalues	3.395	1.580	1.220	1.004	0.874	0.698
Cumulative portion	0.377	0.553	0.688	0.800	0.897	0.975

(b)	Stability Parameters					
	Eigenvectors					
Parameter	1	2	3	4	5	6
CSIM	-0.421	0.060	0.102	0.107	0.078	-
CSIL	-0.337	0.097	0.337	0.513	0.150	-
KI	0.356	0.114	0.213	0.358	0.537	-
TTI	0.370	-0.279	-0.131	0.074	0.530	-
STSI	-0.421	0.063	0.102	0.107	0.074	-
DTA	0.032	-0.381	0.806	-0.450	0.040	-
DTH	-0.042	0.763	0.068	-0.495	0.365	-
θ_{e_8}	0.408	0.195	0.077	-0.030	-0.347	-
θ_{e_7}	0.314	0.355	0.377	0.364	-0.375	-
Eigenvalues	5.346	1.226	0.981	0.881	0.395	0.092
Cumulative portion	0.594	0.730	0.839	0.937	0.981	0.991

Table 16 (Concluded)

(c)	Trigger Parameters					
	Eigenvectors					
Parameter	1	2	3	4	5	6
∇^2_P	-0.370	-0.170	0.129	0.013	0.583	-0.098
ζ	-0.374	-0.079	0.182	-0.254	0.520	-0.009
DVA	-0.012	0.353	0.607	0.615	0.068	0.349
LTHA	-0.149	0.576	-0.231	-0.443	0.050	0.606
IDIV	0.101	-0.073	-0.689	0.542	0.384	0.224
IMDIV	0.105	-0.702	0.157	-0.118	-0.092	0.670
W_{TS}	0.583	0.037	0.133	-0.152	0.344	-0.047
MDIV	0.583	0.099	0.110	-0.176	0.326	-0.019
Eigenvalues	2.170	1.079	1.023	0.991	0.938	0.886
Cumulative portion	0.271	0.406	0.534	0.658	0.775	0.886

the first two components those important variables seem to be equivalent potential temperature at 700 and 850 mb and the Total-Totals Index. If we consider all eigenvectors, the top five parameters are $\theta_{e_5} - \theta_{e_8}$, static stability index, θ_{e_8}, Total-Totals Index, and differential thickness (DTH). All stability parameters are highly intercorrelated and there really should not be many dimensions when they are considered together.

Principal components for trigger parameters indicate that the trigger mechanism is difficult to identify from these parameters. The cumulative variance does not reach 50% until the third eigenvector (compared to first for stability and second for moisture) and 90% is not reached until eigenvector seven (not shown in Table 16, p. 76, as we stop at six eigenvectors). Here, important parameters are vertical motion at the top of the surface layer (this includes terrain induced vertical motion), surface divergence of moisture, and integrated moisture divergence from 850 to 300 mb.

Finally, all predictor parameters can be considered together. This case is summarized in Table 17 where only eigenvalues and cumulative variance are shown. With five principal components one could account for 50% of the total variance among all parameters. Seventeen components could account for 92%. Therefore, it seems justifiable to use at least five variables in discriminant models and possibly up to 17. The radar predictor was not included in this analysis.

h. Operational utility

It is rather fortuitous for individual weather station application

Table 17. Eigenvalues and cumulative portion of total variance accounted for by each successive eigenvector.

	Eigenvector								
	1	2	3	4	5	6	7	8	9
Eigenvalue	9.677	3.032	2.698	2.055	2.012	1.777	1.336	1.218	1.149
Cumulative Portion	0.276	0.363	0.440	0.499	0.556	0.607	0.645	0.680	0.713
	10	11	12	13	14	15	16	17	18
Eigenvalue	1.026	0.997	0.965	0.918	0.872	0.864	0.761	0.734	0.653
Cumulative Portion	0.742	0.771	0.798	0.825	0.849	0.874	0.896	0.917	0.936

that none of the more complicated (from a computational standpoint) parameters were chosen among the top few predictors. In a five-variable equation one would have only to evaluate the moisture divergence term. In order to do this, one needs to plot 1800 GMT mixing ratios obtained from a skew-T diagram along with u and v wind components. A forecaster should extract values of (1) the product of u x W at two east-west grid points spaced 130 km apart, 65 km to either side of his station, and (2) v x W at two similarly spaced north-south grid points.

Negative predicted probabilities are possible but should be considered as zero. Similarly, probabilities greater than one should be interpreted as one. The probability threshold for a thunderstorm-no-thunderstorm decision could be estimated from the 40% postagreement percentages in the contingency tables from within the dependent or total samples. The best estimate for either the total, north wind, or south wind equations is about 0.28. This would optimize prefigurance at the expense of "crying wolf" and total percent correct. Of course, this cut-off would be shifted toward lower probabilities when a critical

(in terms of costs involved) task was involved.

The probabilities can be used directly and the operator should be encouraged to use these in conjunction with cost analyses. If the costs of protective action and loss potential for inaction are known, then the occurrence probabilities can be used in objective cost-loss algorithms (Murphy, 1976).

Operational equations should be developed in given areas with all data available. Since this study was undertaken, an additional year of data has been collected. New equations should incorporate all days for which predictor-predictand samples are available and should be applied to the subsequent year. So long as a few (five or six) predictors are used, weather station forecasters within the development area and for the particular predictor-predictand times could use the equations directly for estimating the probability of thunderstorms. More complicated equations which incorporate extensive analysis, and transformed predictors would be applied to current data at facilities with computer processing capability. Probabilities could be transmitted to appropriate locations. This latter procedure is currently employed by the National Weather Service. (See National Weather Service Technical Procedures Bulletin 194.)

The following five-variable equations developed from the 1974-1975 sample can be tested with current data and probabilities evaluated:

Total
$$PY = 0.0181 + 0.0185*W + 0.414*MDRP - 0.00278*(\vec{V}\cdot\vec{WV}) - 0.00569*(T-T_d)_7 - 0.00515*(\theta_{e_7} - \theta_{e_8}) \quad (8)$$

North wind
$$PY = 1.028 + 0.00337*W + 0.358*MDRP - 0.00336*(\vec{V}\cdot\vec{WV}) - 0.00374*(T-T_d)_7 - 0.00373*\theta_{e_7} \quad (9)$$

$$\text{South wind PY} = 0.0655 + 0.427*\text{MDRP} + 0.0194*W - 0.00265*(\vec{V}\cdot\vec{WV}) - 0.00583*(T-T_d)_8 - 0.00406*(T-T_d)_7 \quad (10)$$

Coefficients from these equations are valid for the following units of measure for predictors: W (g kg^{-1}); MDRP (zero or one for no precip or precip); $\vec{V}$ (m s^{-1})· T, θ, T_d (K); $\Delta x = \Delta y = 1.3$ (m) in the moisture divergence calculation. Predicted probabilities would apply to locations within the developmental area (Fig. 4, p. 32) and are valid with 1800 GMT surface or 1200 GMT upper-air observations. Thunderstorm probabilities (PY) would apply to the area shown in Fig. 6 (p. 39) with respect to the forecasting station and during the period 2000 to 2300 GMT.

Performance in terms of prefigurance and postagreement percentages of a binary (yes or no) forecast could be expected to be slightly lower than the 65%, 40% obtained, respectively, with equations containing more than 15 predictors.

6. UPPER-AIR CONDITIONS AT 3-h INTERVALS

On one day, 24 April 1975, upper-air data were available at 3-h intervals. These were collected as part of the Fourth Atmospheric Variability Experiment (AVE IV) sponsored by the National Aeronautics and Space Administration (NASA). Analyzed fields of temperature, height, dew point, and wind components from a 158-km grid spacing for 49 grid points and four levels were utilized in a test to determine changes of correlations and predictors at different times with occurrences of thunderstorms at 2000-2300 GMT. Analysis procedures are described by Fuelberg (1976).

Twenty-one candidate predictors were calculated for each grid point at 1200 GMT, 1500 GMT, and 1800 GMT. The predictand was the highest MDR value (converted to binary) in an area equivalent to a 138-km box surrounding the grid point as in previous work and for any time during the period 2000 to 2300 GMT. Again, variable selection techniques were used to choose subsets of predictors. Stepwise procedures provided the first several predictors; all possible regressions were considered in the selection of variables four through six. A stepdown or backward elimination procedure was used for those models beyond six variables. Separate regression analyses were performed for each period, and the same candidate predictors as discussed earlier were available to each. Maximum R^2 achieved for each model from a one-variable model up to a model with all 21 variables is shown in Fig. 12. As expected, most of the explained variance was obtained with the first three variables. What is surprising is that the 1800 GMT predictor time, which is closest to the time for which the forecast is made, did not provide a clearly

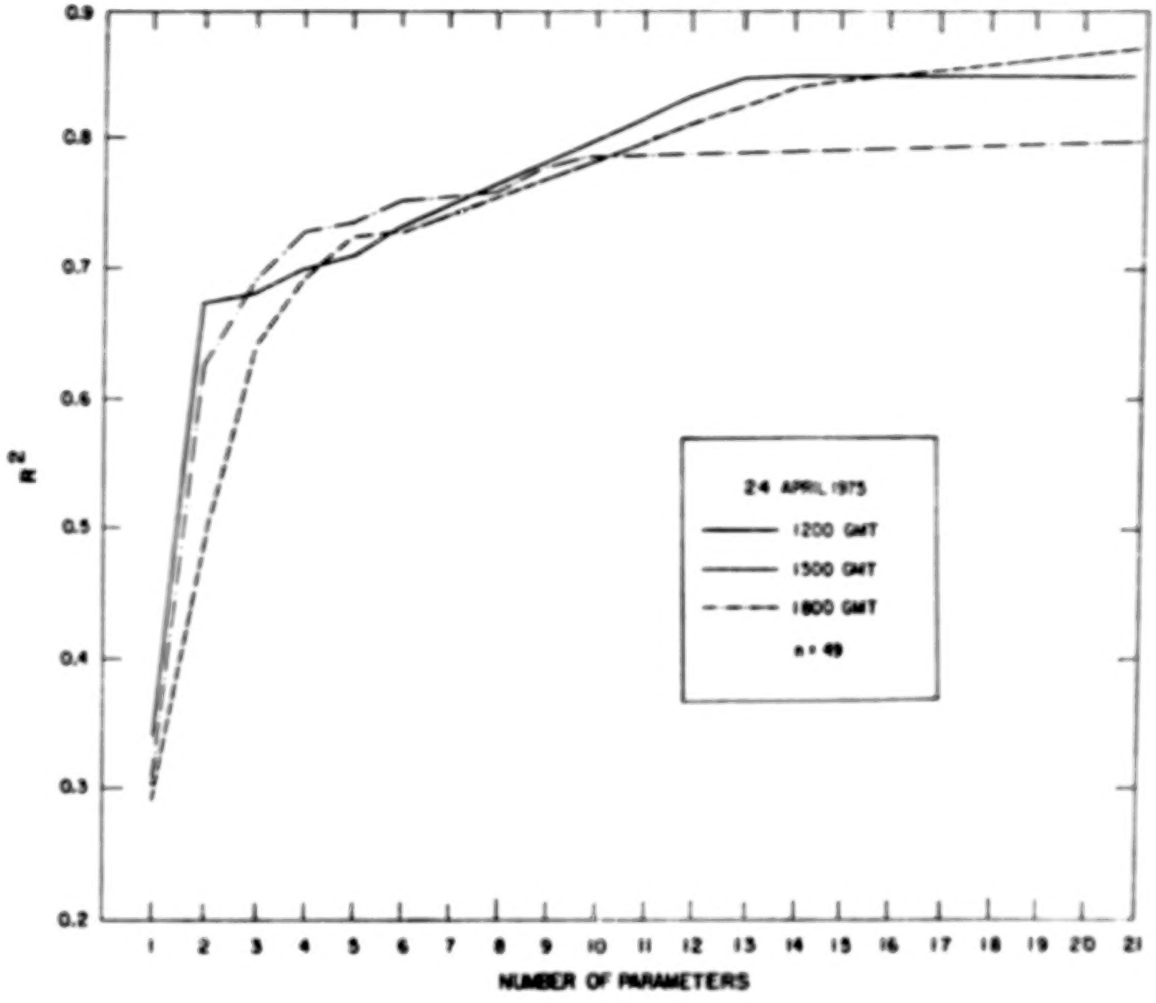


Fig. 12. Fractional amount of total variance in thunderstorm occurrence accounted for by numbers of predictors and a combination of selection procedures.

superior equation. With models including the leading one and two predictors, R^2 is highest for 1200 GMT and lowest for 1500 GMT though the differences are slight. Maximum R^2 of 0.874 was achieved for the all-variable model with 1500 GMT data, whereas the maximum R^2 seems to reach a plateau beyond ten variables for the 1800 GMT period. No completely satisfying explanation is apparent for the lack of improvement as the predictand time is approached; however, there are some

possibilities. As pointed out in Section 2, the assumptions inherent in an analysis of this type are not fulfilled. These errors may be preventing the measure of true correlations. Secondly, this was one day for which there were only 49 observations, and many of these were not independent.

On this day most of the thunderstorm activity was associated with two squall lines. As shown in Fig. 13, the first group of cells was dissipating and moving southeastward between 1200 and 1500 GMT. At 1800 GMT there were few echoes. The second line became active after 2100 GMT. One may hypothesize that there were different atmospheric environments created by the occurrence or nonoccurrence of convection at many of the 49 points for each time. Similarly, a discontinuity existed across the area in the form of a stationary front shown in Fig. 14. Such a feature complicates the interpretation of results for all points as each is considered an independent, separate observation. For example, temperature may be important to thunderstorm development in the area behind (in the cool air) the front, but its influence may be masked by the many observations in the warm air where it may not be important at all. Finally, the response of the atmosphere to the synoptic-scale parameters is being measured. There may be different response times for different parameters. It is possible that those upper-air features at 1800 GMT to which the atmosphere responds most exist on a horizontal and vertical scale smaller than can be resolved from our data.

Table 18 contains the predictors selected during each of the three periods. Up to the five-variable model all antecedent predictors are

(a) 1200 GMT

(b) 1500 GMT

Fig. 13. MDR data for the AVE IV experiment. The code is explained in Table 6.

(c) 1800 GMT

(d) 2100 GMT

Fig. 13. (Concluded)

TABLE OF CONTENTS

		Page	
LIST OF FIGURES		vii	1/A5
LIST OF TABLES		viii	1/A6
LIST OF SYMBOLS		x	1/A8
1.	INTRODUCTION	1	1/A13
a.	Statement of the problem	1	1/A13
b.	Previous studies	2	1/A14
c.	Objectives	7	1/B5
d.	Importance	9	1/B7
2.	STATISTICAL APPROACH	10	1/B8
a.	Linear models	10	1/B8
b.	Partitioning sums of squares	11	1/B9
c.	Model assumptions and violations	13	1/B11
d.	Multicollinearity	19	1/C3
e.	Variable selection methods and inference	22	1/C6
f.	Interpretation of regression model results	24	1/C8
g.	Principal component analysis	28	1/C12
3.	DATA SELECTION AND PROCESSING	31	1/D1
a.	Location	31	1/D1
b.	Surface data	31	1/D1

TABLE OF CONTENTS (CONTINUED)

		Page	
c.	Upper-air data	31	1/D1
d.	Radar data	31	1/D1
e.	Initial processing	35	1/D5
f.	Objective analysis	36	1/D6
4.	PARAMETERIZATION AND DATA SUBDIVISION	38	1/D8
a.	Predictand formulation	38	1/D8
b.	Predictor formulation	38	1/D8
c.	Subdivisions of original data	48	1/E4
5.	RESULTS	52	1/E8
a.	Forecast time intervals	52	1/E8
b.	Predictor selections	53	1/E9
c.	Importance of surface versus upper-air parameters	56	1/E12
d.	Quality of fit of the regression model	60	1/F2
e.	Performance on a test data sample	64	1/F6
f.	Comparison with other results	70	1/F12
g.	Dimensionality	74	1/G2
h.	Operational utility	78	1/G8
6.	UPPER-AIR CONDITIONS AT 3-h INTERVALS	82	1/G10
7.	SUMMARY AND CONCLUSIONS	91	2/A14

LIST OF FIGURES

Figure		Page	
1	Plot of mixing ratio versus residual from regression model	16	1/B14
2	Plot of residuals versus predicted values from regression model	18	1/C2
3	A logistic response function	20	1/C4
4	Data reporting locations	32	1/D2
5	Manually digitized radar grid	34	1/D4
6	Orientation of the predictand area with respect to the predictor point	39	1/D9
7	Schematic illustration of isoline concentration by (a) shear and (b) confluence	45	1/E1
8	Subdivision of total data set	49	1/E5
9	Relation between predicted and observed probabilities of thunderstorm occurrences for various data subdivisions	63	1/F5
10	Regression lines for data in Tables 3 and 4	69	1/F11
11	Frequency distributions of occurrence and nonoccurrence of MDR≥4 for different values of the surface mixing ratio	71	1/F13
12	Fractional amount of total variance in thunderstorm occurrence accounted for by numbers of predictors and a combination of selection procedures	83	1/G11
13	MDR data for the AVE IV experiment	85	1/G13
14	Synoptic charts for 2100 GMT, 24 April 1975 (Fucik and Turner, 1975)	87	2/A10

TABLE OF CONTENTS (Concluded)

	Page	
8. SUGGESTIONS FOR FURTHER RESEARCH	94	2/B3
REFERENCES	97	2/B6
APPENDICES	103	2/B12

LIST OF TABLES

Table		Page	
1	Analysis of variance	12	1/B10
2	Data and regression analysis for 10 observations of hypothetical variables x and y with 10% occurrence frequency	26	1/C10
3	Data and regression analysis for 10 observations of hypothetical variables x and y with 30% occurrence frequency	27	1/C11
4	Data and regression analysis for 57% of nonoccurrence observations in Table 3 data	27	1/C11
5	Comparison of Tables 2, 3, and 4	28	1/C12
6	Explanation of Manually Digitized Radar (MDR) code . . .	33	1/D3
7	Summary of analysis parameters	37	1/D7
8	Candidate predictors	41	1/D11
9	Variables selected, cumulative R^2, and sign of coefficient for a stepwise selection procedure and different data subsets	54	1/E10
10	Linear correlation coefficients of selected predictors with the occurrence of thunderstorms during the period 2000-2300 GMT	57	1/E13
11	Summary of statistics for regression analyses with surface and upper-air predictors	59	1/F1
12	Summary of statistics for regression analyses with different data subsets	61	1/F3
13	Contingency tables for 25 no-thunderstorm forecasts shifted to the yes forecast column in different observed proportions	65	1/F7
14	Summary of contingency tables	67	1/F9
15	Contingency table of observed and forecast thunderstorms for 14 base weather stations near the area outlined in Fig. 4	73	1/G1

LIST OF TABLES (Concluded)

Table		Page	
16	Eigenvectors and eigenvalues for moisture, stability and trigger parameters	76	1/G4
17	Eigenvalues and cumulative portion of total variance accounted for by each successive eigenvector	79	1/G7
18	Summary of AVE IV results for stepwise (A), maximum R^2 reduction (B), and stepdown (C) variable selection techniques	88	2/A11

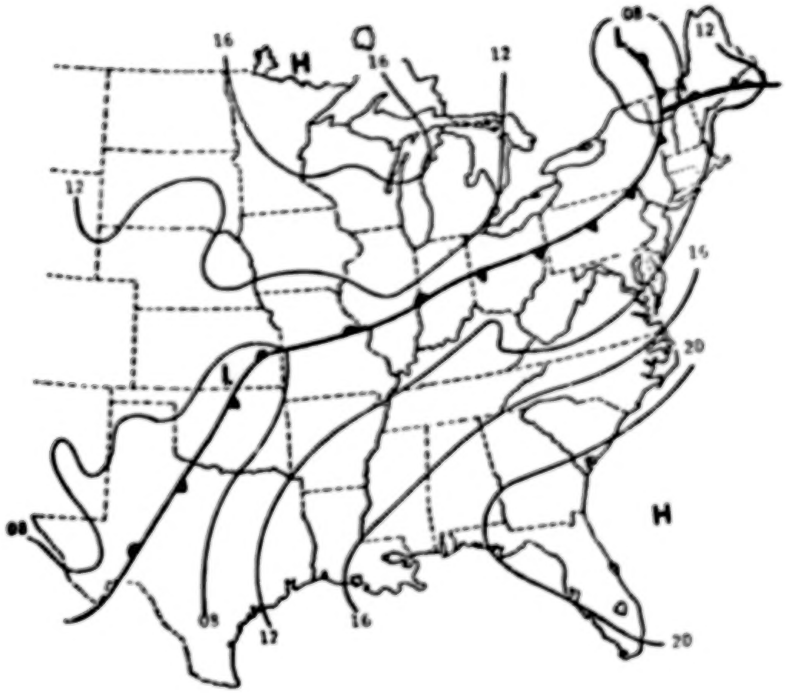
LIST OF SYMBOLS

A_8, A_7, A_5, A_3	subscript convention; A is any variable; A_8, A_7, A_5, A_3 represent the value of A at the 850, 700, 500, and 300 mb pressure levels, respectively
$\lvert\vec{A}\rvert$	magnitude of a two-dimensional vector quantity $\sqrt{A_x^2 + A_y^2}$
ALTSTG	altimeter setting
C_p	Specific heat at constant pressure (1.00488×10^3 J $kg^{-1}K^{-1}$)
CSIL	convective stability index, low level
CSIM	convective stability index, mid level
$\lvert\vec{D}\rvert$	magnitude of a discontinuity function (see text)
DTA	differential temperature advection
DTH	differential thickness
DVA	differential vorticity advection
e	vapor pressure
F	a test statistic for a regression analysis
g	the acceleration of gravity (9.8 m s^{-2})
IDIV	integrated divergence 850 to 500 mb
IMDIV	integrated moisture divergence 850 to 300 mb
KI	K stability index
L	latent heat of condensation (assumed constant at 2.5122×10^6 J kg^{-1})
m	either the number of predictors or when used as a subscript, the mth predictor
MDIV	moisture divergence
MDR	manually digitized radar data
MDRP	manually digitized radar predictor

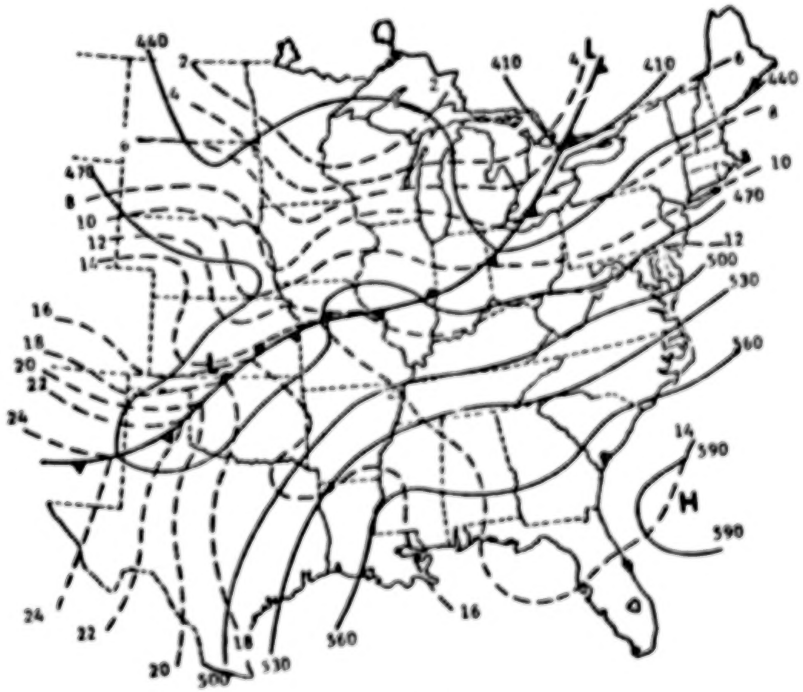
LIST OF SYMBOLS (CONTINUED)

MSE	mean square error
MSR	mean square due to regression
n	total number of observations
P	sea level pressure
p	probability of occurrence
PY	predicted value of Y
R	gas constant for dry air (2.8704×10^2 J $kg^{-1}K^{-1}$)
R^2	ratio of SSR/SST; also called coefficient of variation
SS	skill score
SSE	sum of squares of errors in a regression analysis of variance
SSR	sum of squares due to regression
SST	total corrected sum of squares
STSI	static stability index
T	temperature
T_d	dew point temperature
TS	threat score
TTI	Total Totals Stability Index
u	East-West wind component
UWSH	850 to 500 mb wind shear of the u-wind components
$\vec{V}$	horizontal wind vector
v	North-South wind component
$\lvert\vec{V}_5\rvert$	magnitude of 500 mb wind speed
VSUM	sum of v-wind components at 850 and 500 mb

W	mixing ratio
W_s	saturated mixing ratio
x_{im}	the ith observation for the mth candidate predictor
Y	the dependent variable; the occurrence (one) or nonoccurrence (zero) of thunderstorms
Z	geopotential height
β	partial regression coefficient
$\vec{\nabla}$	two dimensional del operator with vertical component excluded
$\vec{\nabla}^2$	horizontal Laplacian operator $(\partial^2/\partial x^2 + \partial^2/\partial y^2)$
$\vec{\nabla}_p$	two-dimensional del operator on a constant pressure surface
$\vec{\nabla}^2_{THA}$	Laplacian of thickness advection
ε	residual or error term in a regression model
Θ	potential temperature
Θ_e	equivalent potential temperature
Θ_{eA}	advection of equivalent potential temperature
ζ	vertical component of wind curl (vorticity)
ρ	air density
ω_T	terrain induced vertical motion
ω_{TS}	vertical motion at the top of the surface layer (see text)



(a) Surface.



(b) 850 mb.

Fig. 14. Synoptic charts for 2100 GMT, 24 April 1975. (Fucik and Turner, 1975).

Table 18. Summary of AVE IV results for stepwise (A), maximum R^2 reduction (B), and stepdown (C) variable selection techniques.

Number of Variables	Selection Technique	1200 GMT Parameter	1200 GMT R^2	1500 GMT Parameter	1500 GMT R^2	1800 GMT Parameter	1800 GMT R^2
1	A	UWSH	0.341	UWSH	0.287	DTA	0.302
2	A	DTA	0.677	DTA	0.496	DTH	0.628
3	A	STSI	0.681	$\lvert\vec{V}_5\rvert$	0.641	Θ_{e7}	0.678
4	B	CSIM	0.700	DTH	0.696	UWSH	0.729
5	B	$(T-T_d)_7$	0.710	VSUM	0.721	IMDIV	0.739
6	B	UWSH, IMDIV, $\lvert\vec{V}_5\rvert$, W_8, $(T-T_d)_7$, VSUM	0.735	CSIM, TTI, UWSH, DTA, $\lvert\vec{V}_5\rvert$, DTH	0.734	CSIM, TTI, STSI, UWSH, DTA, $\lvert\vec{V}_5\rvert$	0.750
8	C	-	-	-	-	-	-
10	C	-	-	-	-	-	0.763
12	C	-	0.835	-	-	-	0.778
14	C	-	0.849	-	0.842	-	0.796
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
21 (All)	C	-	0.856	-	0.874	-	0.800

included. For example, the best five-variable model with 1200 GMT data includes u-component wind shear, differential temperature advection, static stability index, mid-level convective instability, and the dew-point depression at 700 mb. After five predictors different variables are chosen, some of which were not selected up to that point. Again, the particular variables selected beyond five should not really be discussed since these are undoubtedly more a function of the particular selection technique than any physical mechanism.

The first few variables included in the model can be discussed in that these variables in linear combination are most highly correlated to subsequent thunderstorms on 24 April 1975. The difference between the u-wind component at 500 and 850 mb is important at the earlier two times. This term is related to the mean horizontal temperature gradient in the layer between 850 and 500 mb insofar as the winds are geostrophic. Differential temperature advection between 850 and 500 mb is also an important term as it is among the top two predictors for all times. Temperature advection probably was an important mechanism for creating the instability on this day. It is interesting to note that the u-component wind shear was the first variable selected for the model at both 1200 and 1500 GMT observation times, whereas it is fourth at 1800 GMT. This may be a consequence of the environmental influence of thunderstorms present at the earlier times but almost totally absent at 1800 GMT. From an energy study of this day, Fuelberg (1976) found strong conversion of potential to kinetic energy associated with intensifying convection. The maximum conversion was at 400 mb. A selection of different variables measured at different times or the same variables

in different order could also be a result of differences in atmospheric response to dynamic as opposed to thermodynamic parameters. More work needs to be done in this area.

In summary, the linear combination of upper-air parameters computed from variables measured 2 to 5 h before the predictand time on 24 April 1975 did not explain more of the variance of thunderstorm occurrence at 2000 to 2300 GMT than those measured 8 to 13 h before. Also, there were differences in parameters selected at the different times. Differential temperature advection was important at all times. The vertical wind shear of the east-west wind component was less important to subsequent intense convection when the former was computed from 1800 GMT measurements compared to this parameter measured at 1200 or 1500 GMT. These results may be a consequence of environmental influences of convection at the earlier two times, since little convection was apparent at 1800 GMT. They also might result from violations of model assumptions.

7. SUMMARY AND CONCLUSIONS

Surface, upper-air, and radar observations analyzed onto a 65-km grid were used exclusively to develop equations which relate predictors to subsequent thunderstorms by classical statistical and parameterization techniques. Particular attention was devoted to minimizing errors which result from violations of model assumptions. Raw data were processed to preserve as much detail as can be justified from the original spacing of observing stations. Every fourth point from a 16 x 16 array was included to reduce the spatial correlation naturally present in meteorological data. Variable selection techniques, plots of model residuals, and principal component analyses were used to reduce the multicollinearity present among independent variables. Finally, several different statistical procedures were used to cross-check and confirm results.

Specific synoptic parameters believed to be related to intense convection were calculated from analyses at 1200, 1700, and 1800 GMT and used as candidate predictors in a stepwise variable-selection procedure. Surface and upper-air data were tested separately. The predictand was the occurrence or nonoccurrence of an MDR code of four or greater (assumed to represent thunderstorms) in an area of about 8500 km^2 surrounding a grid point during three subsequent time combinations. The best time was the period from 2000 to 2300 GMT so that only this combination was used in further analyses.

The equations were found to be stable[11] when applied to test data. Also, they contained reasonable parameters as predictors and produced results in contingency tables comparable with present, subjective techniques and with other statistical procedures. Predicted values from developmental and test samples represented actual thunderstorm frequencies of occurrence. This technique can be used to forecast thunderstorms in an operational environment. Furthermore, thunderstorms can be predicted with greater success with this scheme when the surface wind has a northerly component at 1800 GMT.

While not impressive alone, upper-air data seemed to add an important ingredient, namely stability, which is not available from surface data. Radar echoes present at and before the forecast time also added an important dimension. MDR code greater than one near 1700 GMT can lead to MDR of four or greater between 2000 and 2300 GMT due to diurnal effects, or a high MDR initially might tend to persist in space and time. In any case, this radar predictor indicates the presence of vertical motion, a recognized trigger mechanism. Neither time nor space derivatives as computed in this study were particularly important predictors with the notable exception of moisture divergence. But the surface mixing ratio, occurrence of antecedent precipitation, convergence of moisture, and stability were chosen to be among the top five predictors in every case. A reason for the poor showing of other

---

[11]Stable in this context means that statistics in both the developmental and test data sample are nearly the same.

derivatives was that the small-scale gradients important to intense convection cannot be measured due to data-resolution constraints from fixed observation networks.

It was found from both the stepwise procedure and principal-component analyses that linear equations should include from five to 17 variables when parameters represent observed surface and upper-air features. Furthermore, measures of the trigger mechanism were found to be most difficult to define from data in this study, whereas moisture parameters were easily defined.

Equations with many variables will produce slightly better results in terms of prefigurance and postagreement discriminates. Reasonable values to expect would be 65% and 40%, respectively.

Finally, parameters from upper-air observations at 1800 GMT on 24 April 1974 were not more highly correlated to thunderstorms in the period 2000-2300 GMT than were parameters from observations at 1200 or 1500 GMT. This result may be a consequence of the small statistical sample, violations of assumptions in the statistical analysis and the organized development and movement of two groups of thunderstorms. One group influenced observations from which parameters were calculated at 1200 and 1500 GMT.

8. SUGGESTIONS FOR FURTHER RESEARCH

In a study of this scope and magnitude there are practical restrictions on the amount of data to be handled, numbers of predictors used, and types of processing to be performed. It is believed that this research remained within these constraints without sacrificing scientific thoroughness and accuracy. Nevertheless, these limitations and results of the investigation itself provide several suggestions for future research.

a). In order to capture some of the true mesoscale features of the atmosphere, synchronous meteorological satellite data should be used. Mesoscale wind fields determined from satellite cloud observations might be important predictors of severe weather (Houghton and Wilson, 1975). Time and space derivatives of equivalent black body temperatures might reveal small-scale features which lead to subsequent thunderstorms. A microwave sensor, such as that flown on the NASA satellites, would provide indications of soil moisture. Albedo might be important as well. Some preliminary experiments with regression procedures and the ATS-3 satellite data by Sikula and Vonder Haar (1972) indicated satisfactory results when the dependent variables were ceilings and visibilities and independent variables were satellite radiances. Even conventional data available from several mesoscale networks such as HIPLEX (Scoggins and Wilson, 1976), NSSL (Fankhauser, 1969), and METROMEX (Changnon et al., 1971) could be used in this type of study to determine what additional information about subsequent thunderstorms is available for a few areas. Several thunderstorm seasons must be used, however.

b). Severe thunderstorms might be predicted from statistical procedures by use of upper-air winds inferred from satellite thickness (and geopotential height) calculations. Areas of jet streams and difluence aloft could be identified and related to severe weather. Digital radar data now available at several locations (Muench, 1976) could be used as well as additive data from present MDR reports in conjunction with severe weather prediction.

c). Different predictors from conventional data could be tested. For example, present weather, past weather, visibility, wind gusts, sky conditions and remarks are available from surface observations. Climatological frequencies of occurrence for thunderstorms could be computed from all available thunderstorm data and these used as predictors as well. Use of upper-air data should be expanded to include all the resolution in the present observation. In addition, time changes for upper air parameters might be tested. Trajectories of key parameters might make important predictors. The K Index and TTI could both be updated by using the temperature and moisture from 1800 GMT surface observations averaged with those observed at 850 mb 12 h earlier.

d). The area for predictor selection should be allowed to vary and predictand area reduced. The reduction in correlation due to reduced size of predictand might be compensated for by parameters from smaller-scale data sources selected from different areas.

e). More work on the timeliness of upper-air data is required. Additional days when 3-h data are available should be used to obtain a more adequate sample. Similarly, further research into the time changes of surface and upper-air reports should be performed to determine

atmospheric response times (in terms of producing intense convection) for various physical processes such as differential advections.

f). Further work on air mass stratifications would be fruitful. One might use combinations of temperature, wind and moisture to identify three or four types of air masses. Five years of digital radar data will be available for this type of work after the 1977 season.

g). We should continue to investigate random sampling or other ways of reducing the many nonoccurrence days. A forecaster is not concerned with predicting thunderstorms on the many days that he is confident there will be none.

h). There should be more investigation into verification techniques for this type of data.

REFERENCES

Alaka, M. A., W. D. Bonner, J. P. Charba, R. L. Crisci, R. C. Elvander, R. M. Reap, 1973: Objective techniques for forecasting thunderstorms and severe weather. Final Report to Department of Transportation, Federal Aviation Administration, Report Number FAA-RD-73-117, 97 pp.

Auer, August H., Jr., 1976: Observations of an industrial cumulus. J. Appl. Meteor., 15, 406-413.

Barnes, Stanley L., 1964: A technique for maximizing details in numerical weather map analysis. J. Appl. Meteor., 3, 396-409.

_____, 1976: Severe local storms: concepts and understanding. Bull. Amer. Meteor. Soc., 57, 412-423.

Barr, Anthony J., James H. Goodnight, John P. Sall, and Jane T. Helwig, 1976: A User's Guide to SAS 76. SAS Institute, Inc., Sparks Press, Raleigh, N. C., 329 pp.

Beers, Norman R., 1945: Atmospheric stability and instability. Handbook of Meteorology. Berry, F. A., Bollay, E., and Beers, Norman R., editors, McGraw-Hill Book Company, New York, 712-715.

Brandes, Edward A., 1977: Flow in severe thunderstorms observed by dual-doppler radar. Mon. Wea. Rev., 105, 113-120.

Brier, G. W., and R. A. Allen, 1952: Verification of weather forecasts. Compendium of Meteorology, Boston, Amer. Meteor. Soc., 841-848.

_____, and Gayle T. Meltesen, 1976: Eigenvector analysis for prediction of time series. J. Appl. Meteor., 15, 1307-1312.

Browning, K. A., and F. H. Ludlam, 1962: Airflow in convective storms. Quart. J. Roy. Meteor. Soc., 88, 117-135.

Byers, H. R., and R. Braham, Jr., 1949: The Thunderstorm. U.S. Dept. of Commerce, Washington, D.C., 287 pp.

Changnon, S. A., Jr., F. A. Huff, and R. G. Semonin, 1971: METROMEX: An investigation of inadvertent weather modification. Bull. Amer. Meteor. Soc., 52, 958-967.

Charba, J. P., 1975: Operational scheme for short range forecasts of severe local weather. Preprints, Ninth Conf. on Severe Local Storms, Norman, Ok., Amer. Meteor. Soc., 51-57.

_____, 1977: Operational system for predicting thunderstorms two to six hours in advance. NOAA Technical Memorandum NWS TDL-64, Silver Spring, Md., 24 pp.

Dobryshman, E. M., 1972: Review of forecast verification techniques. WMO Tech. Note, No. 120, 17-20.

Donaldson, Ralph J., Jr., Rosemary M. Dyer, and Michael J. Kraus, 1975: An objective evaluator of techniques for predicting severe weather events. Preprints, Ninth Conf. on Severe Local Storms, Boston, Mass., Amer. Meteor. Soc., 321-326.

Draper, N. R., and H. Smith, 1966: Applied Regression Analyses. John Wiley & Sons, New York, 407 pp.

Ellrod, Gary P., and John D. Marwitz, 1976: Structure and interaction in the subcloud region of thunderstorms. J. Appl. Meteor., 15, 1083-1091.

Endlich, R. M., and R. L. Mancuso, 1968: Objective analysis of environmental conditions associated with severe thunderstorms and tornadoes. Mon. Wea. Rev., 96, 342-350.

Essenwanger, Oskar M., 1976: Applied Statistics in Atmospheric Science Chapter 4, Calculation of eigenvalues and eigenvectors. Elsevier Scientific Publishing Company, New York, New York, 348-381.

Fankhauser, J. C., 1969: Convective processes resolved by a mesoscale rawinsonde network. J. Appl. Meteor., 8, 778-798.

_____, 1974: Subcloud air mass and moisture flux attending a Northeast Colorado thunderstorm complex. Preprints, Conf. on Cloud Physics, Tuscon, Ariz., Amer. Meteor. Soc., 271-276.

Fawcett, Edwin B., 1977: Current capabilities in prediction at the National Weather Service's National Meteorological Center. Bull. Amer. Met. Soc., 58, 143-149.

Foster, D. S., and R. M. Reap, 1973: Archiving of Manually-digitized radar data. Techniques Development Laboratory Office Note 73-6, National Weather Service, Silver Springs, Md., 12 pp.

Fucik, N. F., and R. E. Turner, 1975: Data for NASA's AVE IV experiment: 25-mb sounding data and synoptic charts. NASA TMX-64952, George C. Marshall Space Flight Center, Alabama, 458 pp.

Fuelberg, Henry Ernest, 1976: Atmospheric energetics in regions of intense convective activity. Ph.D. Dissertation, Department of Meteorology, Texas A&M University, College Station, Tx., 137 pp.

Fujita, Theodore T., and Horace R. Byers, 1977: Spearhead echo and downburst in the crash of an airliner. Mon. Wea. Rev., 105, 129-146.

Glahn, H. R., and Dale A. Lowry, 1972: The use of Model Output Statistics (MOS) in objective weather forecasting. J. Appl. Meteor., 11, 1203-1211.

Glahn, Harry R., and Joseph R. Bocchieri, 1975: Objective estimation of the conditional probability of frozen precipitation. Mon. Wea. Rev., 103, 3-15.

Global Atmospheric Research Programme, 1972: Parameterization of sub-grid scale processes. GARP Publication Series No. 8, WMO-ICSU Joint Organizing Committee, 4-5.

Harris, Richard J., 1975: A Primer of Multivariate Statistics. New York, Academic Press, 332 pp.

Henz, John F., 1974: Synoptic influences on convective cloud development and precipitation production. Preprints, Conf. on Cloud Physics, Amer. Meteor. Soc., 442-449.

Holton, J. R., 1972: An Introduction to Dynamic Meteorology. Academic Press, New York, N. Y., 112-115.

Houghton, D. D., and T. A. Wilson, 1975: Mesoscale wind fields for a severe storm situation determined from synchronous meteorological satellite (SMS) cloud observations. Preprints, Ninth Conf. on Severe Local Storms, Norman, Ok., Amer. Meteor. Soc., 187-192.

Howcroft, J., and A. Desmarais, 1971: The limited area fine mesh (LFM) model. NWS Technical Procedures Bull., No. 67, 11 pp.

Koch, Steven E., 1975: Mesoscale influences upon the intensity of new cells in two severe local storms. Preprints, Ninth Conf. on Severe Local Storms, Norman, Ok., Amer. Meteor. Soc., 105-112.

Kropfli, R. A., and L. J. Miller, 1976: Kinematic structure and flux quantities in a convective storm from dual-doppler radar observations. J. Atmos. Sci., 33, 520-529.

Lemon, Leslie R., 1976: The flanking line, a severe thunderstorm intensification source. J. Atmos. Sci., 33, 686-694.

Lorenz, E. N., 1956: Emperical orthogonal functions and statistical weather prediction. Scientific Report No. 1, Contract AF 19(604)-1566, Department of Meteorology, MIT, 49 pp.

Ludlam, F. H., 1963: Severe local storms: a review. Meteorological Monographs, 5, 1-31.

Miller, R. C., 1972: Notes on the analysis and severe storm forecasting procedures of the Air Force Global Weather Central. Air Weather Service Technical Report 200 (REV), Scott AFB, IL, 102 pp.

Miller, Sanford R., and Clarence L. David, 1971: A statistically generated aid for forecasting severe thunderstorms and tornadoes. Preprints, Seventh Conf. on Severe Local Storms, Kansas City, Mo. Amer. Meteor. Soc., 42-44.

Mogil, M. H., 1974: Evaluation of severe weather and thunderstorms from forecasts using manually-digitized radar data and the SELS severe weather log. Preprints, Fifth Conf. on Forecasting and Analysis, St. Louis, Mo., Amer. Meteor. Soc., 270-275.

Moore, P. L., A. D. Cummings, and D. L. Smith, 1974: The National Weather Service manually-digitized radar program and its application to precipitation probability forecasting. Preprints, Fifth Conf. on Forecasting and Analysis, St. Louis, Mo., Amer. Meteor. Soc., 69-74.

Morrison, D. F., 1976: Multivariate Statistical Methods. New York, McGraw-Hill, 247-265.

Muench, H. S., 1976: Use of digital radar data in severe weather forecasting. Bull. Amer. Meteor. Soc., 57, 298-303.

Murphy, Allan H., 1976: Decision-making models in the cost-loss ratio situation and measures of the value of probability forecasts. Mon. Wea. Rev., 104, 1058-1065.

Neter, John, and William Wasserman, 1974: Applied Linear Statistical Models. Richard D. Irwin, Inc., Homeward, IL, 842 pp.

Paine, Douglas A., and Michael L. Kaplan, 1974: The linking of multi-scaled energy sources leading to atmospheric development. Scientific Report #1, National Science Foundation Grant No. GA-35250, 12-14.

Palmén, E., and C. W. Newton, 1969: Atmospheric circulation systems. Academic Press, New York, 603 pp.

Panofsky, Hans A., and Glenn W. Brier, 1958: Some Applications of Statistics to Meteorology. Pennsylvania State University, University Park, PA, 95.

Petterssen, S., 1956: Weather Analysis and Forecasting, Vol. I. New York, McGraw-Hill, 200-205.

Purdom, James F. W., 1974: Satellite imagery applied to the mesoscale surface analysis and forecast. Preprints, Fifth Conf. on Weather Forecasting and Analysis, St. Louis, Mo., Amer. Meteor. Soc., 63-68.

_____, 1975: personal communication.

Raymond, David J., 1976: Wave-CISK and convective mesosystems. J. Atmos. Sci., 32, 2392-2398.

Saunders, Frederick, and Robert J. Paine, 1975: The structure and thermodynamics of an intense mesoscale convective storm in Oklahoma. J. Atmos. Sci., 32, 1563-1579.

Schaefer, Joseph T., 1975: Nonlinear biconstituent diffusion: A possible trigger of convection. J. Atmos. Sci., 32, 2278-2284.

Scoggins, James R., 1976: Diurnal and seasonal variations in mesoscale systems. Preprints, Conf. on Meteorology over the Gulf of Mexico, Texas A&M University, College Station, Tx., 43-53.

_____, and Gregory S. Wilson, 1976: Texas HIPLEX mesoscale experiment summer 1976 data tabulations. Final Report, Texas Water Development Board Contract No. 14-60025, Austin, Texas, 1-2.

Scorer, R. S., and F. H. Ludlam, 1953: Bubble theory of penetrative convection. Quart. J. Roy. Meteor. Soc., 79, 94-103.

Shenk, William E., Edward Puccinelli, and Cornelius J. Callahan, 1976: Geosynchronous meteorological satellite data seminar, NASA X-931-76-87, National Aeronautics and Space Administration, Goddard Space Flight Center, Greenbelt, Md., 85-90.

Sikula, G. T., and A. T. Vonder Haar, 1972: Very short range local area weather forecasting using measurements from geosynchronous meteorological satellites. Atmospheric Science Paper #185, Colorado State University, 73 pp.

Smith, W. L., and H. M. Woolf, 1976: The use of eigenvectors of statistical covariance matrices for interpreting satellite sounding radiometer observations. J. Atmos. Sci., 33, 1127-1140.

Squires, P., and J. S. Turner, 1962: An entraining jet model for cumulo-nimbus updraughts. Tellus, 14, 422-434.

Turner, J. S., 1964: The flow into an expanding spherical vortex. J. Fluid Mech., 18, 195-208.

Weiss, Carl E., and James F. W. Purdom, 1974: The effect of early-morning cloudiness on squall-line activity. Mon. Wea. Rev , 102, 400-402.

Whitney, Linwood F., Jr., 1977: Relationship of the subtropical jet stream to severe local storms. Mon. Wea. Rev., 105, 398-412.

Woodcock, Frank, 1976: The evaluation of Yes/No forecasts for scientific and administrative purposes. Mon. Wea. Rev., 104, 1209-1214.

Woodward, Betsy, 1959: The motion in and around isolated thermals. Quart. J. Roy. Meteor. Soc., 85, 144-155.

BLANK PAGE

BLANK PAGE

APPENDICES

APPENDIX A

ANOVA for selected regressions

(1) Total Equations (7 predictors)

Source	Degrees of freedom	Sum of Squares	Mean Square	F Value	R^2
Model	7	239.94357522	34.27765360	294.91	0.22481796
Error	7118	827.33582417	0.11623150		
Corrected Total	7125	1067.27939938			

Parameter	Units	$\underline{\beta}$ estimate	Standard error
Intercept	-	3.16740081	-
MDIV	$gg^{-1}s^{-1} \times 10^{8}$	-0.00276198	0.00020761
W	gg^{-1}	25.43248796	1.64332634
MDRP	(1 or 0)	0.39082300	0.01794391
CSIL	K	-0.00512986	0.00069915
θ_{e_7}	K	-0.01096999	0.00097647
W_7	$gg^{-1} \times 10^{3}$	0.05462544	0.00395796

(2) Northwind equation (7 predictors)

Source	Degrees of freedom	Sum of Squares	Mean Square	F Value	R^2
Model	7	78.07938810	11.15419830	127.84	0.274014
Error	2371	206.86722812	0.08724894		
Corrected Total	2378	284.94661623			

Parameter	Units	$\underline{\beta}$ estimate	Standard error
Intercept	-	1.19350345	0.30434664
MDIV	$gg^{-1}s^{-1} \times 10^{8}$	-0.00303301	0.00034990
$(T-T_d)_7$	K	-0.00430722	0.00084002
W	gg^{-1}	69.09513072	6.42057583
MDRP	0 or 1	0.36068872	0.02911004
T	K	-0.01807394	0.00359192
$T-T_d$	K	0.02365490	0.00360711
θ_{e_7}	K	-0.00485763	0.00101146

(3) Random Equation (6 predictors)

Source	Degrees of freedom	Sum of Squares	Mean Square	F Value	R^2
Model	6	161.91260308	26.98543385	155.31	0.291224
Error	2268	394.05926505	0.17374747		
Corrected Total	2274	555.97186813			

Parameter	Units	$\underline{\beta}$ estimate	Standard error
Intercept	-	1.57191166	0.46848081
MDIV	$gg^{-1}s^{-1} \times 10^{8}$	-0.00340850	0.00039607
$(T-T_d)_7$	K	-0.00789130	0.00129024
W	gg^{-1}	46.44806192	3.21385797
TTI	K	0.01232299	0.00144379
θ_{e_7}	K	-0.00650511	0.00150835
MDRP	0 or 1	0.24007083	0.02659094

(4) Total Equation (20 predictors)

Source	Degrees of freedom	Sum of Squares	Mean Square	F Value	R^2
Model	20	259.37113846	12.96855692	114.05	0.24302084
Error	7105	807.90826092	0.11370982		
Corrected Total	7125	1067.27939938			

Parameter	Units	$\underline{\beta}$ estimate	Standard error
Intercept	-	8.91002397	-
Θ_e	K	0.00762649	0.00105337
MDIV	$gg^{-1}s^{-1} \times 10^{8}$	-0.00228346	0.00023036
$\Theta_e A$	$Ks^{-1} \times 10^{6}$	0.00006796	0.00002784
$\hat{\nabla}^2 P$	$mb\ m^{-2} \times 10^{12}$	0.00008997	0.00005411
ζ	$s^{-1} \times 10^{6}$	0.00036562	0.00023066
$\hat{\nabla}^2 W$	$gg^{-1}m^{-2} \times 10^{15}$	0.00009605	0.00002797
KI	K	0.01872137	0.00316708
DTH	$M\ mb^{-1} \times 10^{2}$	-0.00127867	0.00055703
$(T-T_d)_7$	K	0.01376441	0.00341045
$(T-T_d)_8$	K	0.02245995	0.00355782
v_5	m	0.00262488	0.00048169
W	gg^{-1}	10.97990075	3.09197109
MDRP	(0 or 1)	0.39041505	0.01793046
v	$m\ s^{-1}$	-0.00397307	0.00112285
CSIM	K	-0.15500734	0.02892617

(4) Total Equation (20 predictors) (Concluded)

Parameter	Units	$\underline{\beta}$ estimate	Standard error
STSI	$m^2 sg^{-1}$	0.01596813	0.00284876
Θ_{e_8}	K	-0.03357575	0.00484179
Θ_{e_7}	K	-0.00633453	0.00142566
W_8	$gg^{-1} \times 10^3$	0.14166814	0.01732946
u	$m\ s^{-1}$	-0.00565405	0.00146472

APPENDIX B

Contingency tables for different predicted probability thresholds and dependent equations applied as indicated.

(1) Total dependent equation applied to:

Cut off	Total independent data		Total dependent data	
0.22				
	Yes	No		
Yes	349	187		
No	522	1712		
0.25				
	Yes	No	Yes	No
Yes	314	222	559	247
No	399	1835	708	3216
0.28				
	Yes	No		
Yes	275	261		
No	285	1949		
0.30				
	Yes	No	Yes	No
Yes	243	293	439	367
No	232	2002	419	3505
0.32				
			Yes	No
			Yes	400	406
			No	335	3589

(2) North wind dependent equation applied to:

Cut off	Northwind independent data			Northwind dependent data		
0.25						
		Yes	No		Yes	No
	Yes	80	42	Yes	161	57
	No	126	638	No	200	1200
0.28						
		Yes	No			
	Yes	78	44			
	No	104	660			
0.30						
		Yes	No		Yes	No
	Yes	72	50	Yes	143	75
	No	92	672	No	140	1260
0.34						
		Yes	No		Yes	No
	Yes	57	65	Yes	128	90
	No	68	696	No	96	1304
0.37						
		Yes	No		Yes	No
	Yes	47	75	Yes	114	104
	No	54	710	No	83	1317

(3) Southwind dependent equation applied to:

Cut off	Southwind independent data			Southwind dependent data		
0.25		Yes	No		Yes	No
	Yes	238	176	Yes	407	181
	No	259	1211	No	521	2003
0.27		Yes	No		Yes	No
	Yes	223	191	Yes	372	216
	No	210	1260	No	425	2099
0.30		Yes	No		Yes	No
	Yes	185	229	Yes	330	258
	No	136	1334	No	311	2213
0.33					Yes	No
				Yes	277	311
				No	226	2298

(4) Random sample dependent equation applied to:

Cut off	Total independent data		
0.42		Yes	No
	Yes	451	85
	No	989	1245
0.50		Yes	No
	Yes	408	128
	No	706	1528
0.65		Yes	No
	Yes	281	255
	No	296	1938

	Random independent data			Random dependent data		
0.46		Yes	No			
	Yes	430	106			
	No	123	227			
0.50	Yes	Yes	No		Yes	No
	Yes	408	128	Yes	682	124
	No	103	262	No	209	395

1. REPORT NO.	2. GOVERNMENT ACCESSION NO.	3. RECIPIENT'S CATALOG NO.
NASA CR-2934		

4. TITLE AND SUBTITLE	5. REPORT DATE
Forecasting Thunderstorms over a 2- to 5-h Period by Statistical Methods	December 1977
	6. PERFORMING ORGANIZATION CODE

7. AUTHOR(S)	8. PERFORMING ORGANIZATION REPORT #
Joseph Allen Zak	M-243

9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. WORK UNIT NO.
Department of Meteorology Texas A&M University College Station, Texas 77843	11. CONTRACT OR GRANT NO. NAS8-31773

12. SPONSORING AGENCY NAME AND ADDRESS	13. TYPE OF REPORT & PERIOD COVERED
National Aeronautics and Space Administration Washington, D. C. 20546	Contractor
	14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

This effort was accomplished under the sponsorship of the Atmospheric Sciences Division, Space Sciences Laboratory, NASA/Marshall Space Flight Center, with Mr. Kelly Hill as the contract monitor.

16. ABSTRACT

Classical statistical techniques, such as multiple regression with variable selection and principal component analysis, were employed to define combinations of parameters from meteorological observations which optimally discriminate between the occurrence and nonoccurrence of thunderstorms. Routine observations of weather elements at five levels in the troposphere during two spring and summer seasons were analyzed objectively onto a 65-km grid which spanned much of the central United States. A thunderstorm occurrence was defined from manually digitized radar (MDR) observations with an MDR code of four or greater as the basis. The binary variable one or zero for occurrence or nonoccurrence, respectively, was the predictand. Parameters which are measures of atmospheric moisture content, stability, and trigger mechanisms were calculated from gridded fields of surface and upper-air observed elements for different times each morning. These parameters were candidate predictors in the variable-selection procedures. Data from all grid points and for each day were pooled in order to provide an adequate sample of thunderstorm observations.

Errors which result from usual assumptions in a regression model were quantitatively analyzed. Multicollinearity was severe but minimized through stepwise and maximum R^2 variable selection techniques. Specification and heteroskedasticity errors which result from the binary nature of the dependent variable were present but did not invalidate the overall results.

The first four variables selected in every case were surface mixing ratio, occurrence of precipitation during the morning, moisture convergence, and a stability measure. These four variables include the synoptic-scale conditions commonly recognized as prerequisites for thunderstorms. The trigger mechanism was most difficult to specify from the data, followed by stability, and then moisture. Additional parameters (up to 17) continued to reduce the total, unexplained variance of thunderstorm occurrence. Time changes in surface parameters were not selected as leading predictors. Upper-air observations added an important ingredient, the stability, which, apparently, could not be inferred adequately from surface measurements alone.

Data were grouped by surface wind component, random sampling, and for a spring and summer month. About one-third of the data was saved for a test of results. Thunderstorms were more predictable between 2000 and 2300 GMT when surface winds had a northerly component at 1800 GMT. Random sampling was a way of reducing the influence of the many observations of no thunderstorms which result from the low climatological frequency of occurrence. Predictors in April reflected the importance of kinematics, while those in July were associated with thermodynamic variables as would be expected from synoptic-scale data.

Finally, regression statistics with the predictand being occurrence of thunderstorms at 2000 to 2300 GMT did not show important differences when upper-air parameters were calculated from observations at 1200, 1500, or 1800 GMT. However, these data were available only on one day, 24 April 1975.

The results from this study are comparable with other objective forecasts and with those produced by weather station forecasters although direct comparisons are difficult to make. This technique can be applied rapidly and effectively in an operational environment at locations within the developmental area. It offers all the advantages of an objective forecast and contains no disadvantages from being tied to specific forecast models.

17. KEY WORDS	18. DISTRIBUTION STATEMENT
	Category 47

19. SECURITY CLASSIF. (of this report)	20. SECURITY CLASSIF. (of this page)	21. NO. OF PAGES	22. PRICE
Unclassified	Unclassified	124	$5.50

* For sale by the National Technical Information Service, Springfield, Virginia 22161

NASA-Langley, 1977

END

4.19.78